Unravelling the plant diversity of the Amazonian *canga* through DNA barcoding

Santelmo Vasconcelos[1], Gisele Nunes[1], Mariana Dias[1], Jamily Lorena[1], Renato Oliveira[1], Talvâne Lima[1], Eder Pires[1], Rafael Valadares[1], Ronnie Alves[1], Maurício Watanabe[1], Daniela Zappi[2], Alice Hiura[1], Mayara Pastore[1], Liziane Vasconcelos[1], Nara Mota[3], Pedro Viana[3], André Gil[3], Andre Simões[4], Vera Imperatriz-Fonseca[1], Raymond Harley[5], Ana Giulietti[1], and Guilherme Oliveira[1]

[1]Instituto Tecnológico Vale Desenvolvimento Sustentável
[2]Universidade de Brasilia
[3]Museu Paraense Emilio Goeldi Campus de Pesquisa
[4]Universidade Estadual de Campinas - Campus Cidade Universitaria Zeferino Vaz
[5]Royal Botanic Gardens Kew

April 19, 2021

Abstract

The canga of the Serra dos Carajás, in Eastern Amazon, is home to a unique open plant community, harbouring several endemic and rare species. Although a complete flora survey has been recently published, scarce to no genetic information is available for most plant species of the ironstone outcrops of the Serra dos Carajás. In this scenario, DNA barcoding appears as a fast and effective approach to assess the genetic diversity of the Serra dos Carajás flora, considering the growing need for robust biodiversity conservation planning in such an area with industrial mining activities. Thus, after testing eight different DNA barcode markers (matK, rbcL, rpoB, rpoC1, atpF-atpH, psbK-psbI, trnH-psbA and ITS2), we chose rbcL and ITS2 as the most suitable markers for a broad application in the regional flora. Here we describe DNA barcodes for 1,130 specimens of 538 species, 323 genera and 115 families of vascular plants, with a total of 344 species being barcoded for the first time. In addition, we assessed the potential of using DNA metabarcoding of bulk samples for surveying plant diversity in the canga. Upon achieving the first comprehensive DNA barcoding effort directed to a complete flora in the Brazilian Amazon, we discuss the relevance of our results to guide future conservation measures in the Serra dos Carajás.

Introduction

Conservation efforts depend on a detailed knowledge of the biodiversity in the area of interest, although this is rarely available for megadiverse regions (Myers et al., 2000; Hopkins, 2007; Milliken et al., 2010; Alroy, 2017). The Amazon basin is a vast and diverse biome, being exceptionally important for the maintenance of the biodiversity in the Neotropical region over time (Antonelli et al., 2018). Although the region is undoubtedly one of the most important ecosystems in the planet, harbouring an estimated one quarter of all extant plant species, there is a lack of knowledge about a huge portion of the Amazon ecosystems (Fearnside, 2002; Hopkins, 2007; Milliken et al., 2010; Morim & Lughadha, 2015; BFG, 2018). In addition, along its massive geographic area, the Amazon basin is composed of several different centres of endemism (see Silva et al., 2005 and references within), which are important for the resilience of the forests in face of the disturbing effects of direct anthropological impacts and climate change (Levine et al., 2016).

The Serra dos Carajás (Figure 1), Eastern Amazon, in the southeast of the Brazilian state of Pará, is formed by ironstone outcrops covered by a formation known as *campos rupestres* on *canga* (as detailed in

Souza-Filho et al., 2019; Zappi et al., 2019), surrounded by a dense forest matrix. The *canga* of the Serra dos Carajás is found mostly in the Carajás National Forest (Floresta Nacional de Carajás, or FLONA de Carajás), harbouring several endemic and rare plant species, such as *Philodendron carajasense* E.G.Gonç. (Araceae) and*Carajasia cangae* R.M.Salas, E.L.Cabral & Dessein (Rubiaceae) (Skirycz et al., 2014; Viana et al., 2016; Giulietti et al., 2019), with a high floristic heterogeneity among sites (Zappi et al., 2019). Such ironstone outcrops have been explored throughout the years mainly for iron ore mining activities (Skirycz et al., 2014), and robust biodiversity surveys are necessary to ensure species protection through effective conservation efforts in the presence of industrial activities, especially in view of the climate change scenarios predicted for the region (Levine et al., 2016; Miranda et al., 2019; Giannini et al., 2020).

Plant surveys in the Serra dos Carajás started in the 1970s, as detailed by Viana et al. (2016). However, a project to publish its flora, the Flora of the *canga* of Carajás (FCC), took part in just under four years, being the first complete Flora for a region of the Brazilian Amazon (Mota et al., 2018). This project provided complete floristic treatments for 116 angiosperm families, comprising approximately 900 species (Mota et al., 2018), a number considerably higher than the initial estimate of around 600 species (Viana et al., 2016). Other vascular plant groups detailed in the FCC included 175 ferns in 22 families, 11 lycophytes in three families (Salino et al., 2018), and a single gymnosperm, *Gnetum nodiflorum* Brongn. (Gnetaceae), a liana widely distributed in the Brazilian Amazon (Mota & Giulietti, 2017).

The systematic collection of DNA samples was taken on board as part of the floristic initiative of the FCC project (Mota et al., 2018), as the availability of genetic and genomic data of plants were seen from the onset as extremely important. Such a measure would ensure the correct identification of the species, which had been authenticated by taxonomist specialists, and backed by a deposited voucher, thus guiding more effectively all conservation efforts for the area. The application of DNA barcodes (Hebert et al., 2003) stands out as an efficient source of reliable and cost-effective information for identifying and measuring the diversity status of natural populations of plant species of the*canga* , as recently demonstrated for endemic species such as the morning-glory *Ipomoea cavalcantei* D.F.Austin (Convolvulaceae), and the quillworts *Isoetes cangae* J.B.S.Pereira, Salino & Stützel and *I. serracarajensis* J.B.S.Pereira, Salino & Stützel (Isoetaceae), by Babiychuk et al. (2017) and Nunes et al. (2018), respectively.

However, it is a well-known fact that the development of DNA barcodes is not as straightforward for plants as for other eukaryotes, such as animals and fungi (Fazekas et al., 2009; Hebert et al., 2016; Hollingsworth et al., 2016). The main problems associated with DNA barcoding of plant species arise with the considerably slower pace of evolution of the organelle genomes and the universality of some chloroplast DNA (cpDNA) markers, mainly those with higher nucleotide substitution rates within the plastomes, such as the *mat* K gene (Hollingsworth et al., 2011). Also, there is a difficulty in standardizing which cpDNA regions will function as reliable plant DNA barcodes, since several authors have been reporting variable success rates using different markers (e.g., *rpo* B, *rpo* C1,*atp* F-*atp* H, *psb* K-*psb* I and*trn* H-*psb* A) (e.g., Fazekas et al., 2008), although the combination of the *rbc* L and *mat* K sequences have been recommended as the core barcoding loci (CBOL Plant Working Group, 2009; Kress, 2017). Besides organelle markers, some regions of the nuclear genome, such as the internal transcribed spacers (ITS1 and ITS2) of the 35S rRNA gene, yield useful DNA barcodes for plants (Chen et al., 2010; Hollingsworth et al., 2011).

Furthermore, the generation of DNA barcodes at the species level enables the use of composite samples for detection of species from a given environment, known as DNA metabarcoding. This approach has been regarded as a robust, fast, and cost-effective approach for automated multi-species identification (Deiner et al., 2017; Zinger et al., 2019). For plants, ITS2 has been one of the main markers of choice for surveying multiple species at once, considering the methodological advantages of using this DNA barcode, such as the ease of standardizing PCR conditions and a smaller amplicon size (~450 bp) in comparison with other frequently used regions (Chen et al., 2010; Richardson et al., 2015; Gous et al., 2018). Thus, a curated DNA barcode library and well stablished analytical procedures can provide the basis for the successful application of DNA metabarcoding for monitoring biodiversity (Kress, 2017; Dormontt et al., 2018; Adamowicz et al., 2019).

To the best of our knowledge, there is no other DNA barcoding approach directed to the complete flora of any other region in the Amazon basin. Hence, we describe DNA barcodes for vascular plant species mainly focusing on the *canga* of the Serra dos Carajás, also including plants other from areas in the Brazilian state of Pará that are relevant to an understanding of the biodiversity composition of this mountain range as a whole. We tested the potential of eight commonly used DNA barcode regions and then chose the most suitable markers for a broader application of the DNA barcoding approach in the area, in order to provide robust tools to assess genetic diversity data of the flora of the Amazon basin. Here we followed two main premises: (1) the highest possible marker universality, considering the diversity of taxonomic groups in the *canga* ; and (2) a reasonable standardization and automation of the protocols for sample processing and analyses. Moreover, we aimed to test the potential of DNA metabarcoding analyses with ITS2 for surveying plant diversity in the Serra dos Carajás, taking advantage of the DNA barcode library developed here.

Materials and methods

Plant materials for the DNA barcode procedures

Preferentially, young leaf tissues were sampled for the DNA extractions, although either other vegetative or reproductive structures were employed when needed, as in the case of species of Cactaceae and Eriocaulaceae, for instance. A total of 1,179 specimens of vascular plants from 120 families, 343 genera and 577 species were collected in the Serra dos Carajás and other relevant regions in Eastern Amazon, state of Pará, Brazil (Supplementary Table S1), as part of the FCC project (Viana et al., 2016; Mota et al., 2018; Salino et al., 2018), under ICMBio/MMA permit numbers 47856-2, 48272-6, 53990-1 and 63324-1. Approximately 55% of those samples (645 specimens from 96 families, 243 genera and 370 species) were used to test seven different cpDNA regions (the genes *mat* K, *rbc* L, *rpo* B, and *rpo* C1, and the intergenic spacers *atp* F-*atp* H, *psb* K-*psb* I and *trn* H-*psb* A) and the ITS2 intergenic region (Supplementary Tables S1 and S2). The remaining 534 samples were barcoded only after the selection of the two best markers (*rbc* L and ITS2), as detailed below (Supplementary Tables S1 and S3). The vouchers of all sampled specimens were deposited at the MG herbarium (Museu Paraense Emílio Goeldi, Belém, Pará, Brazil).

Most samples (983, ca. 87%) were collected in 2% CTAB-NaCl saturated buffer, as described by Rogstad (1992), and then stored under refrigeration (~4 °C) until the DNA extraction was carried out. The remaining collected tissues (147, ca. 13%) were dried in silica gel and then stored at room temperature (~25 °C) until processing.

DNA extraction

For the DNA extractions, we established an efficient automated protocol for all plant materials, considering the high diversity of taxonomic groups observed in the *canga* of the Serra dos Carajás. Approximately 20 mg of fresh plant tissue (or ~10 mg for silica dried samples) were separated in 96 racked 1.2 mL collection microtubes (Axygen) with two 3 mm tungsten carbide beads (Qiagen). The samples were frozen in a deep freezer (-80 degC) for 18-20 h and then ground in a TissueLyser II (Qiagen) for 1 min at 30 Hz. Then, 600 μL of extraction buffer (2% w/v CTAB, 0.1 mM Tris-HCl, 20 mM EDTA, 1.4 M NaCl) were added to the ground material and the samples were incubated for 40 min at 60 °C in a water bath. The collection microtubes were centrifuged for 1 min at 4,000 rpm to eliminate debris and 300 μL of the supernatant were transferred to a 96 deep-well U-bottom plate. Afterwards, an automated extraction was performed in a QIAcube HT (Qiagen) with the 'Q protocol V1' of the QIAamp 96 DNA Kit (Qiagen), with minor modifications regarding the sample preparation step, which was carried out without the VXL buffer and including an incubation for 30 s after adding 350 μL of binding buffer ACB, mixing for six times. Also, for some difficult samples, the DNA extractions were performed using the CTAB protocol I described by Weising et al. (2005), with minor modifications (0.5-1.0 g of leaf tissue and 10 mL of the extraction buffer, with the addition of 4% w/v PVP and 0.2% v/v β-mercaptoethanol), followed by the selective precipitation of polysaccharides described by Michaels et al. (1994).

PCR conditions and fragment analysis

We used the same PCR conditions for the seven cpDNA markers (Supplementary Table S4) as follows: 2 µL of genomic DNA extracted in the QIAcube HT robot (or ˜20 ng of genomic DNA from the CTAB protocol I), 1.2 µL of 10× reaction buffer (100 mM Tris-HCl, pH 8.3, and 500 mM KCl), 0.6 µL of 50 mM $MgCl_2$, 1 µL of dNTP mix (2 mM each), 0.25 µL of each primer at 10 µM, 0.5 U of Taq polymerase (Thermo Fisher) and Milli-Q water to a total of 12 µL. For the amplification of ITS2, we added 1 µL of DMSO in the reactions. The PCRs were run in a Veriti 96-Well Thermal Cycler (Applied Biosystems), using the following conditions: initial denaturation at 94 °C for 3 min, followed by 30 cycles of amplification with 1 min at 94 °C, 1 min at 54 °C (except for *mat* K, for which the annealing temperature was 46 °C) and 1 min at 72 °C, with a final extension step at 72 °C for 7 min. Then, the amplified DNA was precipitated with 100 µL of 65% isopropanol for 15 min and centrifuged for 45 min at 4,000 rpm at 10 °C. After discarding the supernatant, 125 µL of cold 70% ethanol were added, and the tubes were centrifuged for 10 min at 4,000 rpm at 10 °C. Finally, the supernatant was discarded again, the tubes were dried at room temperature (ca. 20 °C) for 30 min, and the amplified DNA was resuspended in 10 µL of milli-Q water.

Bidirectional sequencing reactions were carried out using the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems), following a modified version of the manufacturer's protocol. Each reaction consisted of 2 µL of the amplified DNA, 2 µL of 5× sequencing buffer, 0.5 µL of 10 µM primer, 0.5 µL of BigDye Terminator and Milli-Q water to a total of 10 µL. The sequencing reactions were precipitated by adding 2 µL of a 1:1 (v/v) solution containing 3 M sodium acetate (pH 5.2) and 125 mM EDTA (pH 8.0) to each sample, 25 µL of absolute ethanol, briefly vortexing and centrifuging at 4,000 rpm for 40 min at 4 °C. After discarding the supernatant, 35 µL of cold 70% ethanol were added, and the tubes were centrifuged for 10 min at 4 °C. Then, the supernatant was discarded, 10 µL of Hi-Di formamide (Thermo Fisher) were added to the tubes, the samples were denatured for 3 min and then put on ice for at least 2 min, prior to DNA fragment analyses in an ABI 3730 DNA Analyzer (Applied Biosystems).

Sequences assembly, quality control and phylogenetic analyses

We used PIPEBAR (Oliveira et al., 2018) to process all trace files (*.ab1 and *.phd) to generate the assembled consensus of the forward and reverse sequences. Initially, the trace files were converted to fastq files for the quality control step. Converted sequences were then trimmed and filtered using a threshold of 20 PHRED score, also considering a minimum overlap of 15 bp and a minimum similarity of 90% for the assembly step. We defined the "–coding" parameter as 1 and 0 for protein coding genes (*rbc* L, *rpo* B, *rpo* C1 and*mat* K) and intergenic regions (ITS2, *atp* F-*atp* H,*psb* K-*psb* I and *trn* H-*psb* A), respectively. For the protein-coding genes, we defined the "–gcode" parameter as 11 to determine the correct frame for their coding region. Afterwards, to check initially for problematic sequences (from either mislabelled or contaminated samples) generating unusual specimen groupings, considering mainly order and family affiliations, the sequences were aligned with MAFFT 7.388 using the algorithm *Auto* (Katoh & Standley, 2013) for each marker separately. Then, phylogenetic trees based on maximum likelihood (ML) were constructed with RAxML 8.2 (Stamatakis, 2014) as implemented in the CIPRES portal (http://phylo.org), using the substitution model GTR+G and rapid bootstrapping with 1,000 replicates (Supplementary Figures S1-S12). Furthermore, we performed BLASTn searches in the GenBank database (http://blast.ncbi.nlm.nih.gov/Blast.cgi) for additional quality control to avoid problematic sequences, especially in the case of the intergenic regions, which were considerably more difficult to align due to the high taxonomic diversity among the sampled specimens.

Finally, we tested the phylogenetic resolution by counting monophyletic species with at least 70% of bootstrap support, considering only those with more than one sampled specimen. ML trees were constructed in RAxML as described above, using six different matrices based on the*rbc* L and ITS2 alignments, including a topology constraint for the family relationships based on Gastauer & Meira Neto (2017), with minor modifications considering Mota et al. (2018), and PPG I (2016) (Supplementary Data S1; Supplementary Figures S1-S6): two concatenated matrices – (1) *rbc* L+ITS2_full, considering the complete sampling, including accessions with missing sequences of one of the two markers, and (2) *rbc* L+ITS2_cut, considering only specimens with both barcodes; and four matrices from separate alignments – (3)*rbc* L_full and (4) ITS2_full, based on the complete sampling, (5) *rbcL* _cut and (6) ITS2_cut, based on the reduced sampling used in the second matrix.

Barcode analysis

To test the barcode resolution (as the percentage of correctly assigned species) of the eight different markers as barcodes, all-to-all BLAST searches were performed with the sequences obtained herein (both as query and local database), as described in Burgess et al. (2011), using the BLASTn plugin in Geneious Prime 2019.2.3 (Biomatters). Thus, we considered a correct assignment whether a given query sequence presented 100% pairwise identity only with the species itself, in the cases of just one available sequence for the species (336 spp.; 61.9%), such as*Mandevilla tenuifolia* (J.C.Mikan) Woodson (Apocynaceae), or when the intraspecific pairwise identities were either similar or higher when compared with accessions of other species, in the cases of species with more than one specimen with a barcode (207 spp.; 38.1%), such as*Mandevilla scabra* (Hoffmanns. ex Roem. & Schult.) K.Schum. (Supplementary Tables S1 and S3). Additionally, we tested whether the combination of the *rbc* L and ITS2 barcodes (*rbc* L+ITS2) would significantly increase species resolution, following Burgess et al. (2011). Besides, searches were performed in the BOLD database (http://www.boldsystems.org) to check whether there was any barcode previously published for the species analysed in this work, and all sequences produced were deposited in the referred database under the accession numbers listed in the Supplementary Table S1.

Metabarcoding analysis

To assess the potential of using metabarcoding analysis with bulk samples for monitoring plant diversity in the *canga* , we sampled all discernible plant specimens within an approximate 10 m radius in six plots, including two markedly different vegetation types (forest groves and open rupestrian vegetation; Supplementary Table S6), near the end of the dry season (September 27th and 28th, 2017) that lasts from May to October (see Viana et al., 2016). For each sampled locality, pieces of young leaves with approximately 1 cm^2 were collected in a 50 mL Falcon tube containing 30 mL of the 2% CTAB-NaCl saturated buffer, and then stored as previously described.

The procedures for DNA extraction using CTAB and selective precipitation of polysaccharides followed as mentioned above, except for the amounts of leaf tissue (8 g) and extraction buffer (15 mL) per sample. Likewise, the amplification of the ITS2 region followed the same PCR conditions as before, with minor modifications, including 1× TBT-PAR buffer (Samarakoon et al., 2013) and using the primers ITS2-S2F (Chen et al., 2010), with the adapters Ion A, and ITS4 (White et al., 1990), with the adapter trP1. Then, PCR products were purified with the kit Agencourt AMPure XP Beads (Beckman Coulter), following manufacturer's instructions. Each of the six different libraries (one library per collection plot) was composed by pooling four independent PCR replicates and sequenced using the Ion PGM platform (Thermo Fisher).

Raw data from the single-end sequencing run were processed using FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit) and the R package DADA2 (Callahan et al., 2016) to correct sequencing errors and infer exact amplicon sequence variants (ASVs) (equivalent to OTU determination). Firstly, reads were demultiplexed with FASTX Barcode Splitter and sequences of primers and tags were removed with FASTA/Q Clipper. Then, low-quality reads were filtered and trimmed using the filterAndTrim function in DADA2, with maximum expected error of 2 and the following parameters: filterAndTrim, minLen = 90, maxN = 0, maxEE = 2, truncQ = 2, trimLeft = 20, trimRight = 10. The error model was calculated with the function learnErrors based on 100 million randomly selected bases, and reads were dereplicated using derepFastq. Identified chimeric sequences were removed with the removeBimeraDeNovo. An ASVs table was created, and representative sequences were assigned to taxa with BLASTn using our ITS2 library as a local reference database, based on minimum similarity and coverage settings (-perc_identity 95 and -qcov_hsp 70). Finally, we used the LULU curation algorithm with default settings to collapse erroneous ASVs, minimum relative co-occurrence of 0.95 and the default minimum similarity threshold of 84% (Frøslev et al., 2017). Additionally, downstream analyses were performed with the R package Phyloseq v1.26.1 (McMurdie & Holmes, 2013), with an object built from the ASVs curated version, using data from taxonomy assignments and sampling plots.

Results

Amplification and sequencing success of barcodes

Considering only the initial test with the eight markers assessment using 645 samples and 370 species, the proportions of barcoding success (94.73% and 93.24%, respectively) were similar to the complete sampling including the specimens barcoded only with *rbc* L and ITS2. Our results clearly showed *rbc* L (503 samples; 304 species) and ITS2 (490; 286) to be the best barcode regions, with the highest species coverage (Table 1; Supplementary Table S1). On the other hand,*mat* K (154; 126), *rpo* B (136; 119), *rpo* C1 (156; 126),*atp* F-*atp* H (176; 127), *psb* K-*psb* I (125; 95) and *trn* H-*psb* A (109; 77) presented considerably lower numbers of generated sequences, especially these last two regions (Table 1; Supplementary Table S1). Out of the eight species with neither*rbc* L nor ITS2 barcodes, five had sequences of just one of the remaining markers (*Carajasia cangae* , Rubiaceae – *rpo* B;*Sinningia minima* A.O.Araujo & Chautems, Gesneriaceae –*rpo* B; *Myrcia tenuiflora* A.R.Lourenço & E.Lucas, Myrtaceae – *rpo* C1; *Stachytarpheta glabra* Cham., Verbenaceae –*atp* F-*atp* H; and *Hemionitis palmata* L., Pteridaceae – *trn* H-*psb* A), while three had more than one barcode (*Justicia potamogeton* Lindau, Acanthaceae – *mat* K,*rpo* C1 and *atp* F-*atp* H; *Picramnia ferrea* Pirani & W.W.Thomas, Picramniaceae – *rpo* B and *atp* F-*atp* H; and *Senna latifolia* (G.Mey.) H.S.Irwin & Barneby, Fabaceae –*mat* K, *rpo* B and *rpo* C1) (Supplementary Table S1). Additionally, we obtained sequences of all eight barcode markers for only six species: *Aegiphila integrifolia* (Jacq.) Moldenke (Lamiaceae), *Ctenanthe ericae* C.L.Andersson (Marantaceae),*Eriocaulon cinereum* R.Br. (Eriocaulaceae), *Helanthium tenellum* (Mart.) Britton (Alismataceae), *Jacquemontia tamnifolia*(L.) Griseb. (Convolvulaceae) and *Pilocarpus carajaensis* Skorupa (Rutaceae) (Supplementary Tables S1 and S2).

Afterwards, considering the additional 534 samples, we obtained sequences of *rbc* L and ITS2 for other 183 and 140 species (393 and 425 samples), respectively (Supplementary Tables S1 and S3). Almost all barcoded species had, at least, sequences of either *rbc* L or ITS2 (527 out of 535 spp.; 98.50%), from which 399 (75.71%) presented both barcodes (Supplementary Tables S1 and S3).

From our complete sampling, considering the 645 specimens used in the initial test with eight markers, plus the 534 remaining samples barcoded using only *rbc* L and ITS2, we obtained valid sequences of at least one of the eight markers for 538 out of the 575 sampled species (93.56%), totalling 1,130 specimens (95.84%) and 2,729 DNA barcodes (Supplementary Table S1). After searching for previous record in the BOLD database, we observed that 344 (63.94%) of those species were barcoded for the first time in the present work (Supplementary Table S1). In addition, 33 out of the 323 genera with species barcoded here (10.22%) did not have any sequence available in the BOLD database, with several of them being from speciose and representative families in the*canga* , such as Asteraceae (*Cavalcantia* ,*Monogereion* , *Parapiqueria* and *Praxelis*) and Poaceae (*Actinocladum* , *Hildaea* , *Paratheria* ,*Parodiolyra* , *Raddiella* , *Rhytachne* and*Trichanthecium*) (Supplementary Tables S1 and S5). Considering the sampled families, we obtained sequences for 115 out of 120 (95.83%) (Supplementary Table S1).

On the other hand, a total of 49 samples (4.16%) from 37 species (6.43%) could not be barcoded due to problems either with the PCRs or in generating sequencing reads with minimally required quality levels (Supplementary Table S1). Balanophoraceae, Begoniaceae, Dilleniaceae, Hydrocharitaceae and Trigoniaceae were the only families without representatives with a valid barcode sequence (Supplementary Table S1). Samples from some taxa were markedly more challenging to process with the "universal" protocols adopted in this work, and Melastomataceae species were strikingly problematic. Considering that, out of 19 specimens from 13 species and eight genera, only a total of five samples from four species (*Bellucia grossularioides* (L.) Triana,*Miconia heliotropoides* Triana, *Noterophila crassipes*(Naudin.) Kriebel & M.J.R.Rocha and *Tibouchina* sp.) were successfully barcoded (Supplementary Table S1).

It is noteworthy, however, that samples of several species that yielded good quality DNA and amplicons in their expected size ranges presented poor sequencing results for either one or both reads. This problem was more evident in the cases of the cpDNA intergenic regions, for which sequencing results commonly generated electropherograms with many superposed peaks due to the presence of mononucleotide repeats, as observed in reads of *atp* F-*atp* H sequences of *Ipomoea cavalcantei* , for instance (Supplementary Figure S13).

Barcode resolution

Considering the initial test with the eight markers, we observed levels of barcode resolution (percentage of identified species) above 90% for almost all regions, with 97.40% for *trn* H-*psb* A, 94.96% for *rpo* B, 93.65% for *mat* K, 93.70% for*atp* F-*atp* H, 92.66% for ITS2, 92.63% for*psb* K-*psb* I and 90.48% for *rpo* C1 (Table 1). The*rbc* L marker presented the lowest barcode resolution, with 83.22% of the species successfully identified (Table 1). On the other hand, the combined resolution of the two tested markers with the best sequencing results (*rbc* L+ITS2) was higher (92.54%) than the resolution of both regions alone (Table 1).

Considering the complete sampling with *rbc* L and ITS2, that includes the 645 specimens from the test and the 534 additional samples barcoded only with these two regions, the barcode resolution levels were lower, since *rbc* L, ITS2 and *rbc* L+ITS2 presented 75.00%, 89.45% and 86.06%, respectively (Table 2). Moreover, we observed a different barcode resolution pattern when comparing species with only one or with more than one accession. In the case of the species with a single barcoded specimen, the resolution values of both markers (*rbc* L with 77.60% for 317 spp.; and ITS2 with 93.64% for 283 spp.) were considerably higher than for the species with at least two accessions available (*rbc* L with 68.85% for 183 spp.; and ITS2 with 78.21% for 153 spp.) (Table 2; Supplementary Table S3). Likewise, for most genera represented by a single barcoded species, we also observed considerably higher levels of resolution (91.83% and 95.65% for *rbc* L and ITS2, respectively) in comparison with species from the genera with more than one sampled species (67.05% and 86.67% for*rbc* L and ITS2, respectively).

Phylogenetic resolution

Among the phylogenetic trees obtained from the six used matrixes (Figure 2; Supplementary Figures S1-S6), the proportions of species with more than one accession were close, ranging between 34.10-38.87% (Table 2). On the other hand, we observed a wider variation in the percentage of monophyletic species recovered by each matrix, ranging from 62.30% for*rbc* L_full and 79.87% for *rbc* L+ITS2_cut (Table 2; Supplementary Table S3). Considering both complete and reduced matrixes, the ITS2 marker presented higher phylogenetic resolution (75.16% for ITS2_full and 76.12% for ITS2_cut) than *rbc* L (62.30% for*rbc* L_full and 67.91% for *rbc* L _cut) (Table 2). Interestingly, the concatenated matrices presented contrasting patterns of phylogenetic resolution (Table 2). The *rbc* L+ITS2_full matrix presented a considerably lower proportion of monophyletic species (66.02%) than the ITS2_full matrix (75.16%) (Table 2). Conversely, the phylogenetic resolution of *rbc* L+ITS2_cut (79.85%) was higher than both equivalent independent matrices (*rbc* L_cut and ITS2_cut, with 67.91% and 76.12%, respectively) (Table 2).

Additionally, most of the species correctly identified in the barcode resolution analysis were recovered as monophyletic (Supplementary Table S3). Nevertheless, some of the species correctly identified by the DNA barcodes (with barcode resolution) were not resolved in the phylogenies, such as *Clitoria falcata* Lam. (Fabaceae), which was correctly identified in the BLAST analyses with both *rbc* L and ITS2, although appearing as polyphyletic in all six trees (Supplementary Table S3). Correspondingly, the opposite situation, in which the species were monophyletic in all trees but without barcode resolution, was also observed, as in the case of *Lindernia brachyphylla* Pennell (Linderniaceae) (Supplementary Table S3).

Metabarcoding analysis

The ITS2 high-throughput amplicon sequencing generated 4,465,309 raw reads from the composite samples of the six plots (Supplementary Table S6) in the Serra dos Carajás. After the quality control step, 2,269,135 high-quality reads remained, yielding an average length of 314 bp. A total of 508 different ASVs were observed in the metabarcoding analysis after sequence filtering, then being grouped into 41 ASVs classified to the species level, considering 95% and 70% of sequence similarity and coverage, respectively, resulting in 34 identified species, belonging to 33 genera, 21 families and 14 orders (Figure 3; Supplementary Table S7). Malpighiales was the most representative order, with nine species, followed by Asterales, Fabales, Gentianales, Lamiales and Myrtales with three species each (Figure 3; Supplementary Table S7). In general, the distribution of taxa among areas was quite variable, with most observed species being associated with a single collection plot, such as the endemics *Cuphea carajasensis* Lourteig (Lythraceae)*Parapiqueria cavalcantei* R.M.King & H.Rob.

(Asteraceae) and*Perama carajensis* J.H.Kirkbr. (Rubiaceae) (Figure 3; Supplementary Table S7). On the other hand, *Byrsonima stipulacea*A.Juss. (Malpighiaceae), *Croton* sp. (Euphorbiaceae),*Eugenia flavescens* DC. (Myrtaceae), *Forsteronia affinis*Müll.Arg. (Apocynaceae), *Ipomoea marabaensis* D.F.Austin & Secco (Convolvulaceae), *Moquilea egleri* (Prance) Sothers & Prance (Chrysobalanaceae), *Richardia brasiliensis* Gomes (Rubiaceae) and*Sobralia liliastrum* Salzm. ex Lindl. (Orchidaceae) could be identified from samples of at least two different areas.

Discussion

Establishing a reliable DNA barcode library for the flora of the Amazonian campo rupestre *on* canga

The practice of identifying species using DNA sequences is quite old and became mainstream after its formalization by Hebert et al. (2003), as pointed out by DeWalt (2011). The implementation of DNA barcoding approaches for plants was slower and more complex than for animal species (see Fazekas et al., 2009). Nevertheless, the importance of DNA barcodes in surveying plant diversity has been extensively acknowledged during the last decade, despite the inherent difficulties of establishing universal and practical methodologies to be applied in a wide range of taxonomic groups from different ecosystems, considering the particularities observed in several taxa (Hollingsworth et al., 2016; Kress, 2017; Lima et al., 2018). Therefore, we initially tested eight of the most used DNA barcode regions (Fazekas et al., 2008; Lima et al., 2018). This evaluation was important to establish *rbc* L and ITS2 as the best markers covering the principles of standardization, minimalism and scalability (Hollingsworth et al., 2011), considering the vast diversity of vascular plants of the Serra dos Carajás, as recently inventoried by the FCC project (Viana et al., 2016; Mota et al., 2018; Salino et al., 2018).

Considering all vascular plants listed for the *canga* by the FCC project (Viana et al., 2016; Mota et al., 2018; Salino et al.; 2018), our DNA barcodes covered approximately one third of the species diversity (378 out of 1044 spp.; 36.21%). It is important to note that many of the species described for the *canga* are rare and/or difficult to obtain a minimally satisfactory amount of tissue to extract DNA from, some being known only from their type collections, such as the elusive orchid *Uleiorchis longipedicellata* A.Cardoso & Ilk.-Borg. (see Giulietti et al., 2019). On the other hand, sequences for an additional 11 families and 160 species not included in the published lists of the FCC project were obtained, from which 91 were collected in the lowland forest surrounding the *canga* outcrops of the Serra dos Carajás, and 69 from other localities, focusing on the Brazilian state of Pará. Thus, it is essential to emphasize that, despite the fact that the majority of *canga* species still lack barcodes, the total number of species barcoded herein (with 344 out of 538 being barcoded for the first time) characterizes this work as the most extensive DNA barcoding effort for the Brazilian Amazon up to date.

Although defined as one of the two core barcode regions alongside*rbc* L (CBOL, 2009), *mat* K performed poorly in our samples, with amplification and/or sequencing problems in approximately three fourths of the tested specimens. We obtained even worse results for*trn* H-*psb* A, with less than 20% of our samples generating valid sequences, which is surprising since this intergenic region has been one of the preferred alternative barcode markers in several studies (e.g., Lahaye et al., 2008; Erickson et al., 2014). As we have related above, it is paramount to emphasize that many samples were successfully amplified, although with unsatisfactory sequence data recovery, mainly in the cases of the cpDNA intergenic regions. Throughout the history of plant DNA barcoding, there have been several reports of methodological problems with most of the regions tested so far, as frequently observed for *mat* K (e.g., Fazekas et al., 2008; CBOL, 2009; Liu et al., 2015; Ghorbani et al., 2017). On the other hand, the almost fully universal nature of many primers designed to amplify and sequence portions of the *rbc* L gene, obviously including the primer pair we used here, makes this marker the safest choice among the known options to build a comprehensive barcode library for a given flora, even taking into account its lower polymorphism levels among closely related species (Hollingsworth et al., 2011).

Nuclear rDNA-based sequences have been successfully used as DNA barcodes for fungi, especially the ITS region, which is largely employed as the official barcode region for the group (Schoch et al., 2012; Badotti et al., 2017; Wurzbacher et al., 2019). Several authors have emphasized the enormous potential of the ITS components for plant barcoding, which are also frequently regarded as highly informative for resolving phylogenetic relationships (e.g. Liu et al., 2015; Saha et al., 2017; Vasconcelos et al., 2018). Nevertheless, reports of problems with sequence recovery of the complete ITS (including its three regions – ITS1, rDNA 5.8S and ITS2) are not rare for plants, mainly due to issues related to paralogs and pseudogenes (Álvarez & Wendel, 2003; Feliner & Rosselló, 2007). Gonzalez et al. (2009), for instance, obtained poor sequencing results for ITS, with only 41% of the sampled Amazonian trees being successfully barcoded by the authors. On the other hand, the smaller ITS2 region has been indicated as one of the best regions for plant barcoding, presenting a high rate of sequencing success even for lower quality DNA samples (Chen et al., 2010; Kuzmina et al., 2012; Ramalho et al., 2018). Likewise, our data showed the usefulness of ITS2 as the second-best tested marker in terms of sequence recovery, with valid barcodes for 81.04% of the species and 81.33% of the samples and performing relatively close to *rbc* L (91.45% of the species and 79.38% of the samples). Obviously, the availability of sequences of a given marker in public repositories is essential for an effective inventory of plant diversity, and ITS2 has been one of the most frequently used barcode regions for angiosperms so far, accounting for 26.7% of the ca. 340,000 sequences available in the BOLD database (up to January 20th, 2021), only behind *rbc* L and*mat* K, with 35.8% and 31.6%, respectively. Thus, we believe that our choice of implementing ITS2 together with *rbc* L as primary barcodes for this highly diverse flora found in the Amazon basin covers all three principles of DNA barcoding.

Species resolution

Assessing the levels of species discrimination in DNA barcoding approaches is undoubtedly important, although comparing results from different analyses is not as straightforward as one may assume. The first (and perhaps the most important) considerations are related to the study area and sampling coverage. DNA barcoding specific local floras within a well-delimited geographic area, such as the *campo rupestre* on *canga* of the Amazon ironstone fields, for instance, may appear to be more limited in scope than studying the plant diversity of whole countries or broader geographic regions. However, with an original area of ca. 144 km^2 (Souza-Filho et al., 2019), the *canga* of Carajás harbour roughly as many species of vascular plants as Wales, which in fact is a small country, but with an area 152 times larger than the *campo rupestre* on *canga* of Carajás, and with the whole catalogue of 1,143 species of seed plants already barcoded (de Vere et al., 2012). Also, there are basically two main approaches to assess the capability of correctly identifying species (species resolution) of DNA barcode markers. The first is search-based using BLAST (barcode resolution) (e.g., Burgess et al., 2011), and the second one is tree-based, which considers phylogenetic relationships (phylogenetic resolution, tree-based) (e.g., Gonzalez et al., 2009), both with advantages and drawbacks (as discussed below). Therefore, we preferred to use both evaluation approaches.

At first glance, the barcode resolution may seem a more attractive approach, as noticeably higher values were obtained for the two best markers both individually and combined (*rbc* L – 75.00%, ITS2 – 89.45%, and *rbc* L+ITS2 – 86.06%), when compared with the phylogenetic resolution (*rbc* L – 62.30%, ITS2 – 75.16%, and*rbc* L+ITS2 – 66.02%). Moreover, using pairwise identity (or other related parameters of a BLAST search) to determine a correct sequence assignment (and consequently species identification) in DNA barcoding approaches is quite straightforward and practical, especially when handling a large volume of data. On the other hand, the importance of employing a parameter that reflects evolutionary relationships is obvious, as the inclusion of phylogenetic reconstructions with DNA barcoding data enables several other analytical inferences (Erickson et al., 2014; Kress et al., 2015; Miller et al., 2016; Kress, 2017). Therefore, besides assessing the levels of species discrimination of*rbc* L and ITS2 when barcoding the FCC, we also observed indicatives of complex evolutionary relationships either among or within populations of several endemic taxa (as described by Giulietti et al., 2019). In the cases of *Mimosa skinneri* var. *carajarum*Barneby (Fabaceae) and *Borreria elaiosulcata* E.L.Cabral & L.M.Miguel (Rubiaceae), for instance, the trees presented conflicting phylogenetic signals, as non-monophyletic groupings and low bootstrap support were observed (Supplementary Figures S1-S6).

Furthermore, the discrimination levels obtained for both markers (separately and combined) were in accordance with previous results for*rbc* L and ITS2 (e.g., Kress et al., 2009; Burgess et al., 2011; Parmentier et al., 2013), although relatively higher than observed for other diverse floras, as reported by Gonzalez et al. (2009) and Liu et al. (2015). The fact that both species discrimination approaches used here were overly sensitive to sampling coverage is noteworthy, as the analyses considering only specimens with both barcodes provided higher resolution values. This difference was especially strong in the case of the phylogenetic resolution of the combination *rbc* L+ITS2, with an increase of 20.95% in the proportion of resolved species in the reduced sampling in comparison with the complete sampling (from 66.02% to 79.85%; Table 2). Such difference occurred due to the exclusion of specimens from species and/or genera that present either more complex evolutionary histories or problematic taxonomy.

DNA barcodes and conservation

Biodiversity indexes provided by DNA barcoding data have an undeniably important role in better directing conservation efforts, as the effectiveness of maintaining ecological services of biodiversity hotspots can be greatly enhanced by including phylogenetic diversity parameters in the decision-making process (Forest et al., 2007; Diniz et al., 2021). However, as mentioned before and pointed out by Kress (2017), properly populating the public databases with plant DNA barcodes has not been an easy task, being "one of the biggest challenges for the next decade". The difficulties in achieving such an important goal are especially evidenced by considering the actions needed to ensure proper conservation planning in such an immense (and still poorly known) area as the Amazon basin. Hence, it is essential to pay extra attention to endemic and/or rare species of such a unique Amazon vegetation as the*campo rupestre* on *canga* of the Serra dos Carajás, as in the case of the morning-glory *Ipomoea cavalcantei* and the quillwort *Isoetes cangae* , for instance. Both species present a very limited geographic distribution in the *canga* (Giulietti et al., 2019), with studies based on DNA barcoding data investigating their genetic diversity status for the first time (Babiychuk et al., 2017; Nunes et al., 2018), followed by further populational analyses (Lanes et al., 2018; Babiychuk et al., 2019; Dallapicola et al., under review).

Considering the list of endemic plants of the *canga* of Serra dos Carajás, we obtained barcodes for 30 out of the 38 species listed by Giulietti et al. (2019). From the eight endemic species without a DNA barcode, we had access to tissue samples of only two specimens of*Pleroma carajasense* (Melastomataceae), for which we could not obtain DNA sequences of any of the tested markers, as occurred for 47 other samples of 36 species. Likewise, Burgess et al. (2011) had already observed that high throughput DNA isolation procedures would not always work with samples from a wide range of taxonomic groups, with some taxa frequently being more problematic than others, depending on the adopted protocols. The group with the worst performance within our sampling universe was, by far, Melastomataceae, for which we were able to generate sequences only for 26.32% of sampled specimens (and 30.77% of the species). It is important to mention that the Melastomataceae was recorded as the fifth most diverse angiosperm family in the FCC, with a total of 41 species (Rocha et al., 2017; Mota et al., 2018). Lima et al. (2018) also reported low rates of amplification success for *rbc* L and *mat* K when barcoding tree species of Melastomataceae, one of the most species-rich families in the flora of the Brazilian state of São Paulo. Although these authors could overcome such a problem with the plastid markers by using the ITS region, the results we obtained here for the four barcoded species of Melastomataceae with ITS2 were only slightly better than for *rbc* L (one barcoded species), considering the universal protocols used. Thus, we acknowledge the crucial need for developing more directed protocols aiming at problematic taxa, which will be our next step towards accomplishing a DNA barcode library with full coverage for the flora of the Amazonian *canga* .

As mentioned above, inventorying species through DNA-based tools has consistently gained ground along the years, achieving further importance with the development of multispecies identification approaches based on high-throughput sequencing technologies (Kress et al., 2015; Deiner et al., 2017). Several authors have pointed out the many advantages of using DNA metabarcoding for monitoring biodiversity, especially considering robustness and efficiency of this analytical system (Deiner et al., 2017; Zinger et al., 2019; Bush et al., 2020). Certainly, the effectiveness of metabarcoding can be greatly affected depending on the completeness level of the reference DNA barcode library (Alsos et al., 2018), thus, care must be taken for its use for identification of specimens at species level until a complete barcode library is available, especially in areas with several narrowly distributed endemics. Nevertheless, the results obtained here for the bulk samples from Serra dos Carajás were very promising, as we could observe a relatively high taxonomic diversity within and among the collection sites, even with a coverage of less than one third (30.75%) of the*canga* species with ITS2 barcodes. Thus, the validity of DNA metabarcoding with ITS2 for monitoring plant species in Serra dos Carajás was successfully demonstrated, despite having a yet incomplete DNA barcode library.

Furthermore, the value of DNA barcoding data to guide conservation efforts in the Serra dos Carajás has been demonstrated also in the ecological context by helping to identify the importance of some plant taxa acting as nutrient providers for animal communities in ferruginous caves (Ramalho et al., 2018). Besides, plant diversity monitoring projects using metabarcoding approaches with environmental DNA (eDNA) and bulk sampling are currently being established and implemented on a large scale for the plants of Serra dos Carajás (Oliveira et al., 2019). Hence, the development of the DNA barcode libraries for the region will be essential for the optimization of reforestation in decommissioned mining sites in the region, as well as fast and robust vegetation surveys in untouched native areas.

Acknowledgements

This work was funded by Vale (Biodiversity Genomics, RBRS000603.85) and GO is funded by CNPq (Conselho Nacional de Desenvolvimento Científico; 444227/2018-0, 402756/2018-5, 307479/2016-1) and the and the CABANA project (RCUK/BB/P027849/1). DZ (305301/2018-7), GO (307479/2016-1), LVV (380502/2016-0 and 380517/2017-5), MCD (380290/2016-2), MP (443365/2015-6) and TGL (381271/2018-8) were supported by CNPq fellowships. The authors would like to thank the laboratory technician Manoel Lopes and all the taxonomists involved in the PFC initiative.

References

Adamowicz, S. J., Boatwright, J. S., Chain, F., Fisher, B. L., Hogg, I. D., Leese, F., Liftmaer, D. A., Mwale, M., Naaum, A. M., Pochon, X., Steinke, D., Wilson, J. J., Wood, S., Xu, J., Xu, S., Zhou, X., & Van den Bank, M. (2019). Trends in DNA barcoding and metabarcoding.*Genome, 62* , v-vii.

Alroy, J. (2017). Effects of habitat disturbance on tropical forest biodiversity. *Proceedings of the National Academy of Sciences of the United States of America, 114* , 6056-6061.

Alsos, I. G., Lammers, Y., Yoccoz, N. G., Jørgensen, T., Sjögren, P., Gielly, L., & Edwards, M. E. (2018). Plant DNA metabarcoding of lake sediments: How does it represent the contemporary vegetation. *PLoS ONE, 13* , e0195403.

Álvarez, I., & Wendel, J. F. (2003). Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution, 29* , 417-434.

Antonelli, A., Zizka, A., Carvalho, F. A., Scharn, R., Bacon, C.D., Silvestro, D., & Condamine, F. L. (2018). Amazonia is the primary source of Neotropical biodiversity. *Proceedings of the National Academy of Sciences of the United States of America, 115* , 6034-6039.

Babiychuk, E., Kushnir, S., Vasconcelos, S., Dias, M. C., Carvalho-Filho, N., Nunes, G. L., Santos, J. F., Tyski, L., Silva, D. F., Castilho, A., Imperatriz-Fonseca, V. L., & Oliveira, G. (2017). Natural history of the narrow endemics *Ipomoea cavalcantei* and*I. marabaensis* from Amazon Canga savannahs. *Scientific Reports, 7* , 7493.

Babiychuk, E., Teixeira, J. G., Tyski, L., Guimaraes, J. T. F., Romeiro, L. A., Silva, E. F., Santos, J. F., Vasconcelos, S., Silva, D. F., Castilho, A., Siqueira, J. O., Fonseca, V. L. I., & Kushnir, S. (2019). Geography is essential for reproductive isolation between florally diversified morning glory species from Amazon canga savannahs.*Scientific Reports, 9* , 18052.

Badotti, F., Oliveira, F. S., Garcia, C. F., Vaz, A. B. M., Fonseca, P. L. C., Nahum, L. A., Oliveira, G., & Góes-Neto, A. (2017). Effectiveness of ITS and sub-regions as DNA barcode markers for the identification of Basidiomycota (Fungi). *BMC Microbiology, 17* , 42.

BFG, The Brazilian Flora Group (2018) Brazilian Flora 2020: Innovation and collaboration to meet Target 1 of the Global Strategy for Plant Conservation (GSPC). *Rodriguésia, 69* , 1513-1527.

Burgess, K. S., Fazekas, A. J., Kesanakurti, P. R., Graham, S. W., Husband, B. C., Newmaster, S. G., Percy, D. M., Hajibabaei, M., & Barrett., S. C. H. (2011). Discriminating plant species in a local temperate flora using the *rbc* L+*mat* K DNA barcode.*Methods in Ecology and Evolution, 2* , 333-340.

Bush, A., Monk, W. A., Compson, Z. G., Peters, D. L., Porter, T. M., Shokralla, S., Wright, M. T. G., Hajibabaei, M., & Baird, D. J. (2020). DNA metabarcoding reveals metacommunity dynamics in a threatened boreal wetland wilderness. *Proceedings of the National Academy of Sciences of the United States of America, 117* , 8539-8545.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods, 13* , 581-583.

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* , 12794-12797.

Chen, S., Yao, H., Han, J., Liu, C., Song, J., Shi, L., Zhu, Y., Ma, X., Gao, T., Pang, X., Luo, K., Li, Y., Li, X., Jia, X., Lin, Y., & Leon, C. (2010). Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS ONE, 5* , e8613.

Dalapicolla, J., Alves, A., Jaffé, R., Vasconcelos, S., Pires, E. S., Nunes, G. L., Pereira, J. B. S., Guimarães, J. T. F., Dias, M. C., Fernandes, T. N., Scherer, D., Santos, F. M. G., Castilho, A., Santos, M. P., Calderón, E. N., Martins, R. L., Fonseca, R. N., Esteves, F. A., Caldeira, C. F., Oliveira, G. (under review) Conservation implications of genetic structure in the narrowest endemic quillwort from the Eastern Amazon.

de Vere, N., Rich, T. C. G., Ford, C. R., Trinder, S. A., Long, C., Moore, C. W., Satterthwaite, D., Davies, H., Allainguillaume, J., Ronca, S., Tatarinova, T., Garbett, H., Walker, K., & Wilkinson, M. J. (2012). DNA barcoding the native flowering plants and conifers of Wales.*PLoS ONE, 7* , e37945.

Deiner, K., Bik, H. M., Machler, E., Seymour, M., Lacoursiere-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E., & Bernatchez, L. (2017). Environmental DNA metabarcoding: transforming how we survey animal and plant communities.*Molecular Ecology, 26* , 5872-5895.

DeWalt, R. E. (2011). DNA barcoding: a taxonomic point of view.*Journal of the North American Benthological Society, 30* , 174-181.

Diniz, E. S., Gastauer, M., Thiele, J., & Meira-Neto, J. A. A. (2021). Phylogenetic dynamics of Tropical Atlantic Forests. *Evolutionary Ecology, 35* , 65-81.

Dormontt, E. E., van Dijk, K., Bell, K. L., Biffin, E., Breed, M. F., Byrne, M., Caddy-Retalic, S., Encinas-Viso, F., Nevill, P. G., Shapcott, A., Young, J. M., Waycott, M., & Lowe, A. J. (2018). Advancing DNA barcoding and metabarcoding applications for plant requires systematic analysis of herbarium collections – an Australian perspective.*Frontiers in Ecology and Evolution, 6* , 134.

Erickson, D. L., Jones, F. A., Swenson, N. G., Pei, N., Bourg, N. A., Chen, W., Davies, S. J., Ge, X.-J., Hao, Z., Howe, R. W., Huang, C.-L., Larson, A. J., Lum, S. K. Y., Lutz, J. A., Ma, K., Meegaskumbura, M., Mi, X., Parker, J. D., Fang-Sun, I., Wright, S. J., Wolf, A. T., Ye, W., Xing, D., Zimmerman, J. K., & Kress, W. J. (2014) Comparative evolutionary diversity and phylogenetic structure across multiple forest dynamics plots: a mega-phylogeny approach. *Frontiers in genetics, 5* , 358.

Fazekas, A. J., Burgess, K. S., Kesanakurti, P. R., Graham, S. W., Newmaster, S. G., Husband, B. C., Percy, D. M., Hajibabaei, M., & Barrett, S. C. H. (2008). Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS ONE, 3* , e2802.

Fazekas, A. J., Kesanakurti, P. R., Burgess, K. S., Percy, D. M., Graham, S. W., Barrett, S. C., Newmaster, S. G., Hajibabaei, M., & Husband, B. C. (2009). Are plant species inherently harder to discriminate than animal species using DNA barcoding markers?*Molecular Ecology Resources, 9* Suppl s1, 130-139.

Fearnside, P. M. (2002). Biodiversity as an environmental service in Brazil's Amazonian forests: risks, value and conservation.*Environmental Conservation, 26* , 305-321.

Feliner, G. N., & Rosselló, J. A. (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.

Forest, F., Grenyer, R., Rouget, M., Davies, T. J., Cowling, R. M., Faith, D. P., Balmford, A., Manning, J. C., Procheş, Ş., van der Bank, M., Reeves, G., Hedderson, T. A. J., & Savolainen, V. (2007). Preserving the evolutionary potential of floras in biodiversity hotspots. *Nature, 445* , 757-760.

Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A. J. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates.*Nature Communications, 8* , 1188.

Gastauer, M., & Meira Neto, J. A. A. (2017). Updated angiosperm family tree for analyzing phylogenetic diversity and community structure.*Acta Botanica Brasilica, 31* , 191-198.

Ghorbani, A., Gravendeel, B., Selliah, S., Zarre, S., & de Boer, H. (2017). DNA barcoding of tuberous Orchidoideae: a resource for identification of orchids used in Salep. *Molecular Ecology Resources, 17* , 342-352.

Giannini, T. C., Costa, W. F., Borges, R. C., Miranda, L., Costa, C. P. W., Saraiva, A. M., & Imperatriz-Fonseca, V. L. (2020). Climate change in the Eastern Amazon: crop-pollinator and occurrence-restricted bees are potentially more affected. *Regional Environmental Change, 20* , 9.

Giulietti, A. M., Giannini, T. C., Mota, N. F. O., Watanabe, M. T. C., Viana, P. L., Pastore, M., Silva, U. C. S., Siqueira, M. F., Pirani, J. R., Lima, H. C., Pereira, J. B. S., Brito, R. M., Harley, R. M., Siqueira, J. O., & Zappi, D. C. (2019). Edaphic endemism in the Amazon: vascular plants of the canga of Carajás, Brazil. *The Botanical Review, 85* , 357-383.

Gonzalez, M. A., Baraloto, C., Engel, J., Mori, S. A., Pétronelli, P., Riéra, B., Roger, A., Thébaud, C., & Chave, J. (2009). Identification of Amazonian trees with DNA barcodes. *PLoS ONE, 4* , e7483.

Gous, A., Swanevelder, D. Z. H., Eardley, C. D., & Willows-Munro, S. (2019). Plant-pollinator interactions over time: pollen metabarcoding from bees in a historic collection. *Evolutionary Applications, 12* , 187-197.

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

Hebert, P. D. N., Hollingsworth, P. M., & Hajibabaei, M. (2016). From writing to reading the encyclopedia of life. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371* , 20150321.

Hollingsworth, P. M., Graham, S. W., & Little, D. P. (2011). Choosing and using a plant DNA barcode. *PLoS ONE, 6* , e19254.

Hollingsworth, P. M., Li, D.-Z., van der Bank, M., & Twyford, A. D. (2016). Telling plant species apart with DNA: from barcodes to genomes.*Philosophical Transactions of the Royal Society B: Biological Sciences, 371* , 20150338.

Hopkins, M. J. G. (2007). Modelling the known and unknown plant biodiversity of the Amazon Basin. *Journal of Biogeography, 34* , 1400-1411.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability.*Molecular Biology and Evolution, 30* , 772-780.

Kress, W. J. (2017). Plant DNA barcodes: applications today and in the future. Journal of *Systematics and Evolution, 55* , 291-307.

Kress, W. J., Erickson, D. L., Jones, F. A., Swenson, N. G., Perez, R., Sanjur, O., & Bermingham, E. (2009). Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama.*Proceedings of the National Academy of Sciences, 106* , 18621-18626.

Kress, W. J., García-Robledo, C., Uriarte, M., & Erickson, D. L. (2015). DNA barcodes for ecology, evolution, and conservation.*Trends in Ecology & Evolution, 30* , 25-35.

Kuzmina, M. L., Johnson, K. L., Barron, H. R., & Hebert, P. D. N. (2012). Identification of the vascular plants of Churchill, Manitoba, using a DNA barcode library. *BMC Ecology, 12* , 25.

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

Lanes, E. C., Pope, N. S., Alves, R., Carvalho-Filho, N. M., Giannini, T. C., Giulietti, A. M., Imperatriz-Fonseca, V. L., Monteiro, W., Oliveira, G., Silva, A. R., Siqueira, J. O., Souza-Filho, P. W., Vasconcelos, S., & Jaffé, R. (2018). Landscape genomic conservation assessment of a narrow-endemic and a widespread morning glory from Amazonian savannas. *Frontiers in Plant Science, 9* , 532.

Levine, N. M., Zhang, K., Longo, M., Baccini, A., Phillips, O. L., Lewis, S. L., Alvarez-Dávila, E., Andrade, A. C. S., Brienen, R. J. W., Erwin, T. L., Feldpausch, T. R., Mendoza, A. L. M., Vargas, P. N., Prieto, A., Silva-Espejo, J. E., Malhi, Y., & Moorcroft, P. R. (2016). Ecosystem heterogeneity determines the ecological resilience of the Amazon to climate change. *Proceedings of the National Academy of Sciences of the United States of America, 113* , 793-797.

Lima, R. A. F., Oliveira, A. A., Colletta, G. D., Flores, T. B., Coelho, R. L. G., Dias, P., Frey, G. P., Iribar, A., Rodrigues, R. R., Souza, V. C., & Chave J. (2018). Can plant DNA barcoding be implemented in species-rich tropical regions? A perspective from São Paulo State, Brazil. *Genetics and Molecular Biology, 41* , 661-670.

Liu, J., Yan, H.-F., Newmaster, S. G., Pei, N., Ragupathy, S., & Ge, X.-J. (2015). The use of DNA barcoding as a tool for the conservation biogeography of subtropical forests in China. *Diversity and Distributions, 21* , 188-199.

McMurdie, P. J., & Holmes, S. (2013). Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE, 8* , e61217.

Michaels, S. D., John, M. C., & Amasino, R. M. (1994). Removal of polysaccharides from plant DNA by ethanol precipitation.*Biotechniques, 17* , 274-276.

Miller, S. E., Hausmann, A., Hallwachs, W., & Janzen, D. H. (2016). Advancing taxonomy and bioinventories with DNA barcodes.*Philosophical Transactions of the Royal Society B: Biological Sciences, 371* , 20150339.

Milliken, W., Zappi, D., Sasaki, D., Hopkins, M., & Pennington, R. T. (2010). Amazon vegetation: how much don't we know and how much does it matter? *Kew Bulletin, 65* , 691-709.

Miranda, L. S., Imperatriz-Fonseca, V. L., & Giannini, T. C. (2019). Climate change impact on ecosystem functions provided by birds in southeastern Amazonia. *PLoS ONE, 14* , e0215229.

Morim, M. P., Lughadha, E. M. N. (2015). Flora of Brazil Online: Can Brazil's botanists achieve their 2020 vision? *Rodriguésia, 66* , 1115-1135.

Mota, N. F. O., & Giulietti, A. M. (2016). Flora of the cangas of the Serra dos Carajás, Pará, Brazil: Gnetaceae. *Rodriguésia, 67* , 1191-1194.

Mota, N. F. O., Watanabe, M. T. C., Zappi, D. C., Hiura, A. L., Pallos, J., Viveros, R. S., Giulietti, A. M., & Viana, P. L. (2018). Amazon canga: the unique vegetation of Carajás revealed by the list of seed plants. *Rodriguésia, 69* , 1435-1488.

Myers, N., Mittermeier, R. A., Mittermeier, C. G., Fonseca, G. A. B., & Kent, J. (2000). Biodiversity hotspots for conservation priorities.*Nature, 403* , 853-858.

Nunes, G. L., Oliveira, R. R. M., Guimarães, J. T. F., Giulietti, A. M., Caldeira, C., Vasconcelos, S., Pires, E., Dias, M., Watanabe, M., Pereira, J., Jaffé, R., Bandeira, C. H. M. M., Carvalho-Filho, N., Silva, E. F., Rodrigues, T. M., Santos, F. M. G., Fernandes, T., Castilho, A., Souza-Filho, P. W. M., Imperatriz-Fonseca, V., Siqueira, J. O., Alves, R., & Oliveira, G. (2018). Quillworts from the Amazon: a multidisciplinary populational study on *Isoetes serracarajensis*and *Isoetes cangae* . *PLoS ONE, 13* , e0201417.

Oliveira, G., Nunes, G., Valadares, R., Alves, R., & Vasconcelos, S. (2019). DNA barcoding and genomics in the megadiverse Amazon altitude fields. *iBOL Barcode Bulletin, 9* , 5498.

Oliveira, R. R. M., Nunes, G. L., Lima, T. G. L., Oliveira G., & Alves R. (2018). PIPEBAR and OverlapPER: tools for a fast and accurate DNA barcoding analysis and paired-end assembly. *BMC Bioinformatics, 19* , 297.

Parmentier, I., Duminil, J., Kuzmina, M., Philippe, M., Thomas, D. W., Kenfack, D., Chuyong, G. B., Cruaud, C., & Hardy, O. J. (2013). How effective are DNA barcodes in the identification of African rainforest trees? *PLoS ONE, 8* , e54921.

PPG I (2016). A community-derived classification for extant lycophytes and ferns. *Journal of Systematics and Evolution, 54* , 563-603.

Ramalho, A. J., Zappi, D. C., Nunes, G. L., Watanabe, M. T. C., Vasconcelos, S., Dias, M. C., Jaffé, R., Prous, X., Giannini, T. C., Oliveira, G., & Giulietti, A. M. (2018). Blind testing: DNA barcoding sheds light upon the identity of plant fragments as a subsidy for cave conservation. *Frontiers in Plant Science, 9* , 1052.

Richardson, R. T., Lin, C.-H., Sponsler, D. B., Quijia, J. O., Goodell, K., & Johnson, R. M. (2015). Application of ITS2 metabarcoding to determine the provenance of pollen collected by honey bees in an agroecosystem. *Applications in Plant Sciences, 3* , apps.1400066.

Rocha, K. C. J., Goldenberg, R., Meirelles, J., & Viana, P. L. (2017). Flora of the cangas of Serra dos Carajás, Pará, Brazil: Melastomataceae.*Rodriguésia, 68* , 997-1034.

Rogstad, S. H. (1992). Saturated NaCl-CTAB solution as a means of field preservation of leaves for DNA analyses. *Taxon, 41* , 701-708.

Saha, P. S., Sengupta, M., & Jha, S. (2017). Ribosomal DNA ITS1, 5.8S and ITS2 secondary structure, nuclear DNA content and phytochemical analyses reveal distinctive characteristics of four subclades of*Protasparagus* . *Journal of Systematics and Evolution, 55* , 54-70.

Salino, A., Arruda, A. J., & Almeida, T. E. (2018). Ferns and lycophytes from Serra dos Carajás, an Eastern Amazonian mountain range.*Rodriguésia, 69* , 1417-1434.

Samarakoon, T., Wang, S. Y., & Alford, M. H. (2013). Enhancing PCR amplification of DNA from recalcitrant plant specimens using a trehalose-based additive. *Applications in Plant Sciences, 1* , apps.1200236.

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., & Chen, W. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America, 109* , 6241-6246.

Silva, J. M. C., Rylands, A. B., & Fonseca, G. A. B. (2005). The fate of the Amazonian areas of endemism. *Conservation Biology, 19* , 689-694.

Skirycz, A., Castilho, A., Chaparro, C., Carvalho, N., Tzotzos, G., & Siqueira, J.O. (2014). Canga biodiversity, a matter of mining.*Frontiers in Plant Science, 5* , 653.

Souza-Filho, P. W. M., Giannini, T. C., Jaffé, R., Giulietti, A. M., Santos, D. C., Nascimento Jr., W. R., Guimarães, J. T. F., Costa, M. F., Imperatriz-Fonseca, V. L., & Siqueira, J. O. (2019). Mapping and quantification of ferruginous outcrop savannas in the Brazilian Amazon: a challenge for biodiversity conservation. *PLoS ONE, 14* , e0211095.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30* , 1312-1313.

Vasconcelos, S., Soares, M. L., Sakuragui, C. M., Croat, T. B., Oliveira, G., & Benko-Iseppon, A. M. (2018). New insights on the phylogenetic relationships among the traditional *Philodendron*subgenera and the other groups of the *Homalomena* clade (Araceae).*Molecular Phylogenetics and Evolution, 127* , 168-178.

Viana, P. L., Mota, N. F. O., Gil, A. S. B., Salino, A., Zappi, D. C., Harley, R. M., Ilkiu-Borges, A. L., Secco, R. S., Almeida, T. E., Watanabe, M. T. C., Santos, J. U. M., Trovó, M., Maurity, C., & Giulietti, A. M. (2016). Flora of the cangas of the Serra dos Carajás, Pará, Brazil: history, study area and methodology. *Rodriguésia, 67* , 1107-1124.

Weising, K., Nybom, H., Wolff, K., & Kahl, G. (2005). *DNA fingerprinting in plants: Principles, methods, and applications* (2nd ed.). Boca Raton, FL: CRC Press.

White, T. J., Bruns, T., Lee, S., Taylor, J. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: M. A. Innis, D. H. Gelfand, J. J. Sninsky, & T. J. White (Eds.),*PCR Protocols: a guide to methods and applications* (pp. 315-322). San Diego, CA: Academic Press.

Wurzbacher, C., Larsson, E., Bengtsson-Palme, J., Van den Wyngaert, S., Svantesson, S., Kristiansson, E., Kagami, M., & Nilsson, R. H. (2019). Introducing ribosomal tandem repeat barcoding for fungi. *Molecular Ecology Resources, 19* , 118-127.

Zappi, D. C., Moro, M. F., Walker, B., Meagher, T., Viana, P. L., Mota, N. F., Watanabe, M. T. C., & Lughadha, E. N. (2019). Plotting a future for Amazonian canga vegetation in a campo rupestre context. *PLoS ONE, 14* , e0219753.

Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., Chariton, A. A., Creer, S., Coissac, E., Deagle, B. E., De Barba, M., Dickie, I. A., Dumbrell, A. J., Ficetola, G. F., Fierer, N., Fumagalli, L., Gilbert, M. T. P., Jarman, S., Jumpponen, A., Kauserud, H., Orlando, L., Pansu, J., Pawlowski, J., Tedersoo, L., Thomsen, P. F., Willerslev, E., & Taberlet, P. (2019). DNA metabarcoding – need for robust experimental designs to draw sound ecological conclusions.*Molecular Ecology, 28* , 1857-1862.

Data Accessibility

All DNA sequences generated for this work may be accessed through the BOLD accession numbers indicated in the Supplementary Table S1, as well as the sequence alignments provided in the he supplementary data. All other data will be made available upon request.

Author Contributions

The study was conceived and designed by S.V., G.L.N., P.L.V., V.L.I.F., A.M.G. and G.O.; the specimens were either collected or processed by S.V., M.C.D., J.L., R.B.S.V., M.T.C.W., D.C.Z., A.L.H., M.P., L.V.V., N.F.O.M., P.L.V., A.S.B.G, A.O.S., R.M.H and A.M.G.; the laboratory experiments were executed by S.V., M.C.D., J.L. and E.S.P.; the bioinformatic analyses were designed and performed by S.V., G.L.N., J.L., R.R.M.O., T.G.L.L., R.A. and G.O.; the manuscript was written by S.V. and G.L.N. The final version of the manuscript was read and approved by all the authors.

Figure captions

Figure 1. Distribution of the *canga* formation in the Serra dos Carajás, Pará, Brazil. The circumscriptions of the Carajás National Forest (CNF) and Campos Ferruginosos National Park (CFNP) are evidenced within the Amazon Forest.

Figure 2. Maximum likelihood tree from the *rbc* L and ITS2 concatenated matrix of the *canga* plants of the Serra dos Carajás and related regions in the Eastern Amazon. The coloured branches correspond to the listed orders.

Figure 3. Relative abundance of the observed species in the DNA metabarcoding analysis with bulk samples collected in six different*canga* plots in the Serra dos Carajás, as detailed in the Supplementary Table S6.

Supplementary material captions

Supplementary Figure S1. Phylogenetic trees (cladogram in A and phylogram in B) of the *canga* plants of Serra dos Carajás and related regions based on maximum likelihood analysis with *rbc* L sequences. Bootstrap values ([?] 70) are shown above the branches of the cladogram.

Supplementary Figure S2. Phylogenetic trees (cladogram in A and phylogram in B) of the *canga* plants of Serra dos Carajas and related regions based on maximum likelihood analysis with ITS2 sequences. Bootstrap values ([?] 70) are shown above the branches of the cladogram.

Supplementary Figure S3. Phylogenetic trees (cladogram in A and phylogram in B) of the *canga* plants of Serra dos Carajas and related regions based on maximum likelihood analysis with the concatenated matrix with all available *rbc* L and ITS2 sequences (*rbc* L+ITS2_full). Bootstrap values ([?] 70) are shown above the branches of the cladogram.

Supplementary Figure S4. Phylogenetic trees (cladogram in A and phylogram in B) of the *canga* plants of Serra dos Carajas and related regions based on maximum likelihood analysis with the reduced*rbc* L matrix (*rbc* L_cut), including only samples with both*rbc* L and ITS2 sequences available. Bootstrap values ([?] 70) are shown above the branches of the cladogram.

Supplementary Figure S5. Phylogenetic trees (cladogram in A and phylogram in B) of the *canga* plants of Serra dos Carajas and related regions based on maximum likelihood analysis with the reduced ITS2 matrix (ITS2_cut), including only samples with both *rbc* L and ITS2 sequences available. Bootstrap values ([?] 70) are shown above the branches of the cladogram.

Supplementary Figure S6. Phylogenetic trees (cladogram in A and phylogram in B) of the *canga* plants of Serra dos Carajas and related regions based on maximum likelihood analysis with the concatenated matrix of *rbc* L and ITS2 (*rbc* L+ITS2_cut), including only samples with both sequences. Bootstrap values ([?] 70) are shown above the branches of the cladogram.

Supplementary Figure S7. Phylogenetic trees (cladogram in A and phylogram in B) of the *canga* plants of Serra dos Carajas and related regions based on maximum likelihood analysis with *mat* K sequences. Bootstrap values ([?] 70) are shown above the branches of the cladogram.

Supplementary Figure S8. Phylogenetic trees (cladogram in A and phylogram in B) of the *canga* plants of Serra dos Carajas and related regions based on maximum likelihood analysis with *rpo* B sequences. Bootstrap values ([?] 70) are shown above the branches of the cladogram.

Supplementary Figure S9. Phylogenetic trees (cladogram in A and phylogram in B) of the *canga* plants of Serra dos Carajas and related regions based on maximum likelihood analysis with *rpo* C1 sequences. Bootstrap values ([?] 70) are shown above the branches of the cladogram.

Supplementary Figure S10. Phylogenetic trees (cladogram in A and phylogram in B) of the *canga* plants of Serra dos Carajas and related regions based on maximum likelihood analysis with*atp* F-*atp* H sequences. Bootstrap values ([?] 70) are shown above the branches of the cladogram.

Supplementary Figure S11. Phylogenetic trees (cladogram in A and phylogram in B) of the *canga* plants of Serra dos Carajas and related regions based on maximum likelihood analysis with*psb* K-*psb* I sequences. Bootstrap values ([?] 70) are shown above the branches of the cladogram.

Supplementary Figure S12. Phylogenetic trees (cladogram in A and phylogram in B) of the *canga* plants of Serra dos Carajas and related regions based on maximum likelihood analysis with*trn* H-*psb* A sequences. Bootstrap values ([?] 70) are shown above the branches of the cladogram.

Supplementary Figure S13. Representative electropherogram of an*atp* F-*atp* H sequence of an *Ipomoea cavalcantei* sample evidencing several superimposed base call peaks. Red, green, yellow and blue peaks correspond to A, T, G and C, respectively. Base call qualities are represented in boxes coloured in different shades of blue below the peaks, with the lightest and darkest tones being the highest and the lowest qualities, respectively.

Supplementary Table S1. List of all samples used in the DNA barcoding of*canga* plants of the Serra dos Carajas and other related regions, bringing taxonomic information, BOLD accessions, voucher numbers, presence in the species list of the Flora of the *canga* of Carajas (FCC), as described in Mota et al. (2018), whether the sample was included in the test with the eight different markers, presence of previously published DNA barcode for the species in the BOLD database, and availability of the sequences per marker. We obtained sequences of at least one of the markers for all accessions, except for the ones with the names marked in red.

Supplementary Table S2. Barcode resolution analysis for species identification with the eight DNA barcode markers tested (*mat* K,*rbc* L, *rpo* B, *rpo* C1, *atp* F-*atp* H,*psb* K-*psb* I, *trn* H-*psb* A and ITS2) through the BLAST approach.

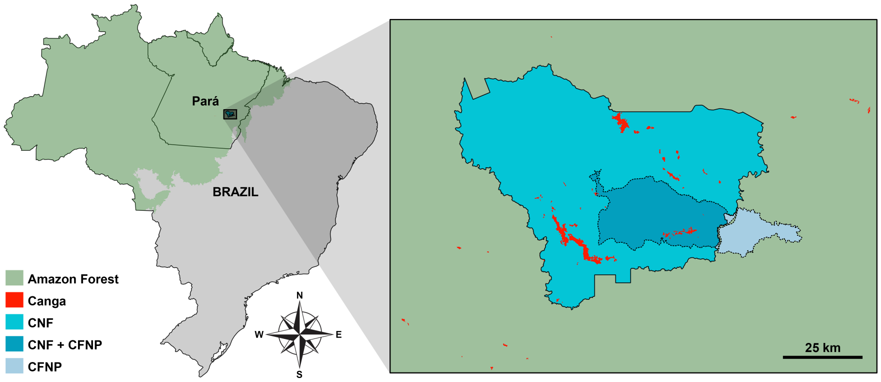
Supplementary Table S3. Barcode resolution (BLAST approach) and phylogenetic resolution (phylogenetic reconstruction approach) of the*canga* plants of the Serra dos Carajas and other related regions with *rbc* L and ITS2. The total number of accessions per species per marker is shown, as well as the different results of both analyses using the six different alignment matrices (*rbc* L_full,*rbc* L_cut, ITS2_full, ITS2_cut, *rbc* L+ITS2_full and*rbc* L+ITS2_cut).

Supplementary Table S4. Primers used to generate DNA barcodes for the plant species of the canga of Serra dos Carajas and related regions in the Brazilian state of Para, Eastern Amazon, as used and referred by Babiychuk et al. (2017).

Supplementary Table S5. Genera without previous DNA barcode records in the BOLD database, indicating their respective families and species sampled in the present work.

Supplementary Table S6. Samples collected for the DNA metabarcoding analysis using bulk samples of collected leaves, indicating vegetation characteristics and the coordinates of the sampling spots.

Supplementary Table S7. Identified ASVs in the DNA metabarcoding analysis with bulk samples of *canga* plant leaves, indicating their taxonomic attribution, BLAST parameters obtained and number of recovered sequences in each of the six different sampling spots, as detailed in Supplementary Table S6.



figures/Figure-2-color-tree/Figure-2-color-tree-eps-converted-to.pdf figures/Figure-3/Figure-3-eps-converted-to.pdf

Hosted file

Table_1.pdf available at https://authorea.com/users/408673/articles/518601-unravelling-the-plant-diversity-of-the-amazonian-canga-through-dna-barcoding

Hosted file

Table_2.pdf available at https://authorea.com/users/408673/articles/518601-unravelling-the-plant-diversity-of-the-amazonian-canga-through-dna-barcoding