Botanical Journal of the Linnean Society, 2018, **188**, 213–231. With 6 figures.

Towards integrative taxonomy in Neotropical botany: disentangling the *Pagamea guianensis* species complex (Rubiaceae)

EDUARDO M. B. PRATA[1]*, CHODON SASS[2], DORIANE P. RODRIGUES[3], FABRICIUS M. C. B. DOMINGOS[4], CHELSEA D. SPECHT[5], GABRIEL DAMASCO[6], CAMILA C. RIBAS[7], PAUL V. A. FINE[6] and ALBERTO VICENTINI[1]

[1]*Coordenação de Dinâmica Ambiental e Programa de Pós-Graduação em Botânica, Instituto Nacional de Pesquisas da Amazônia (INPA), 69060-001, Manaus, AM, Brazil*
[2]*Department of Plant and Microbial Biology, Department of Integrative Biology and the University and Jepson Herbaria, University of California, Berkeley, CA, USA*
[3]*Laboratório de Evolução Aplicada, Universidade Federal do Amazonas (UFAM), 69077-000, Manaus, AM, Brazil*
[4]*Instituto de Ciências Biológicas e da Saúde, Universidade Federal de Mato Grosso, 78698-000, Pontal do Araguaia, MT, Brazil*
[5]*School of Integrative Plant Sciences and the L. H. Bailey Hortorium, Cornell University, 14853, Ithaca, NY, USA*
[6]*Department of Integrative Biology and the University and Jepson Herbaria, University of California, Berkeley, CA, USA*
[7]*Coordenação de Biodiversidade, Instituto Nacional de Pesquisas da Amazônia (INPA), 69060-001, Manaus, AM, Brazil*

Received 19 July 2017; revised 8 June 2018; accepted for publication 23 June 2018

Species complexes are common in the Neotropical flora, and the *Pagamea guianensis* complex is one of the most widespread groups of species in the Amazonian white-sand flora. Previous analyses suggested the occurrence of ten species in this group, but species limits remained unclear due to poor sampling, morphological overlap and low molecular variation. Here we present the most comprehensive population and molecular sampling across the geographical distribution of the *P. guianensis* complex to date in order to test the monophyly of this group and to clarify species limits. Using a high-throughput DNA sequencing approach, we sequenced 431 loci (>34 M bases) for 179 individuals. We applied phylogenetic and species tree analyses to resolve phylogenetic relationships among the sampled individuals. Species delimitation was inferred based on genomic data, and we tested whether hypothesized species could be differentiated using morphological, ecological and near-infrared spectroscopy data. We confirm the monophyly of the *P. guianensis* complex and accept 15 distinct and well-supported lineages, here proposed as 14 species and one subspecies. Our findings highlight the importance of multiple lines of evidence from independent datasets in the process of species delimitation and species discovery in species complexes in the Neotropics.

ADDITIONAL KEYWORDS: Amazonian white-sand flora – coalescent species delimitation – next-generation sequencing – NIR spectroscopy – phylogenetic analysis – species discovery – species tree.

INTRODUCTION

Species complexes are groups in which species limits and hence species numbers are unclear. They usually result from many factors including cryptic morphological variation, poor sampling, large geographical distributions, introgression and recent divergence generating shared ancestral polymorphism (Grube & Kroken, 2000). The taxonomic consequences of cryptic morphological variation in such cases

*Corresponding author. E-mail: eduardombprata@gmail.com

include synonymization or the circumscription of morphologically overlapping species as a single species (a potential false negative for species) and description of new species based on intraspecific morphological variation of an already described species (a potential false positive), the latter resulting in taxonomic inflation (Fujita *et al.*, 2012). Disentangling species limits in species complexes has been one of the main challenges in modern taxonomy, especially since the advent of DNA sequencing methods and the development of phylogenetic and phylogeographic analyses based on molecular data (Grube & Kroken, 2000; Carstens *et al.*, 2012; Carstens, Lemmon & Lemmon, 2013).

During the evolutionary process of species formation, morphological differentiation, habitat specialization and geographical range evolution represent intrinsic factors that can (1) be a consequence of and (2) promote lineage divergence (Mayr, 1992; de Queiroz, 2007). Thus, at any point during the process of species delimitation, a given species may or may not be recognized and circumscribed depending on the criteria and the species concept adopted (de Queiroz, 1999, 2007; Fujita *et al.*, 2012). Historically, many species concepts have been proposed (Mayr, 1992; Van Valen, 1976; Cracraft, 1983; Donoghue, 1985; Mishler & Brandon, 1987; Mayr, 1992) and, despite their particularities, implicitly or explicitly all of them 'equate species with population level evolutionary lineages' (de Queiroz, 1999, 2007). Therefore, because different criteria require different information about divergent lineages, multiple lines of evidence are needed for species delimitation (de Queiroz, 1999, 2007), especially in cases of cryptic speciation and the untangling of species complexes (Costa-Silva *et al.*, 2015; Piedra-Malagón *et al.*, 2016).

Recently, many cryptic species have been discovered based on DNA data (Hebert *et al.*, 2004; Leaché & Fujita, 2010; Garcia *et al.*, 2011; Carstens & Satler, 2013; Domingos *et al.*, 2014; Vicentini, 2016). However, phylogenetic relationships among closely related species are usually difficult to assess because gene trees often conflict with the 'true' species tree as a consequence of incomplete lineage sorting, hybridization and gene duplication leading to potential paralogy (Maddison, 1997; Knowles & Carstens, 2007; Yang & Rannala, 2010; Rannala & Yang, 2015). Thus, the recovered genealogy will depend on the amount of data collected (e.g. the number of genes), the history of the sampled genes and the tree inference method utilized (Chen *et al.*, 2007). High-throughput DNA sequencing methods combined with phylogenetic analyses based on multispecies coalescent models (Heled & Drummond, 2010; Mirarab & Warnow, 2015; Yang, 2015; Edwards *et al.*, 2016) can potentially recover an accurate species tree, even when paraphyletic gene trees are present. These methods are especially useful for disentangling shallow phylogenetic relationships between cryptic species early in the divergence process (Knowles & Carstens, 2007; Yang & Rannala, 2010; Rannala & Yang, 2015).

In the highly diverse Amazon forest (ter Steege *et al.*, 2016), species complexes are commonly present in plant families including Burseraceae (Fine *et al.*, 2005; Fine, Zapata & Daly, 2014), Lauraceae (Vicentini, 1999) and Rubiaceae (Vicentini, 2007, 2016). In the Amazonian white-sand flora, one of the most common genera is *Pagamea* Aubl. (Rubiaceae), represented by shrubs or small trees predominantly found on sand soil habitats, from open savannas to tall forests on dry or flooded soil systems (Vicentini, 2016). A recent phylogenetic analysis based on few (three) molecular markers permitted the recognition of *c*. 30 species in this genus, but the lack of molecular resolution, incongruence among markers, poor sampling and overlapping morphological variation precluded a clear understanding of the limits among some closely-related species (Vicentini, 2007, 2016). Among these, a group of species related to *P. guianensis* Aubl., the *P. guianensis* complex (PGC), includes morphologically similar species among which the limits remain unclear. The PGC includes seven described species: *P. dudleyi* Steyerm., *P. guianensis*, *P. plicatiformis* Steyerm., *P. puberula* Steyerm., *P. pilosa* (Standl.) Steyerm., *P. sessiliflora* Spruce ex Benth., *P. spruceana* Vicent. & E.M.B.Prata and four hypothesized (but not published) species with the informal names *P.* m. *macrocarpa* (called *P.* m. *cryptica* here), *P.* m. *occulta*, *P.* m. *peruviana* and *P.* m. *resinosa* (where 'm.' stands for morphotype, as used in Vicentini, 2016), corresponding to the clades '*Guianensis*' and '*Peruviana*' Vicentini (2016). The geographical distribution of the PGC encompasses the Amazon and the Orinoco river basins, the Guiana and Brazilian Shields and the Atlantic Forest (Vicentini, 2007, 2016). In Vicentini's (2007, 2016) recent review of *Pagamea*, species limits and phylogenetic relationships in the PGC are not clear because of the non-monophyly and morphological overlap among many of these taxa. Thus, the molecular and morphological evidence available is still not sufficient for a clear definition of species limits in the PGC.

Given the broad distribution and high lineage diversity of the PGC, this group provides a great opportunity to apply modern phylogenetic and species delimitation methods to the still understudied Amazonian flora. To achieve this goal, here we implemented the most comprehensive sampling and high-throughput DNA sequencing across the geographical distribution of this group to answer the following questions. (1) Are the PGC and its currently described species monophyletic? (2) How many species does the PGC include and how are these species supported by molecular, morphological, ecological and spectral data? To answer these questions, we first conducted a high-throughput DNA sequencing

protocol and generated a phylogenetic tree to evaluate whether previously described taxa (species, subspecies and the whole complex) are monophyletic. Second, after updating our species hypotheses based on the phylogenetic and species tree analyses, we tested species limits using coalescent species delimitation analysis. Finally, after defining species limits based on molecular data, we analysed whether they are supported by morphological, ecological and near-infrared spectroscopy data.

MATERIAL AND METHODS

Sample collection

Five hundred and seventy-five individual plants of the PGC were collected and prepared as herbarium specimens (Table S1). Flowers and fruits, when present, were preserved in 70% ethanol, and leaf tissue was stored in silica gel for DNA extraction. Our sampling sites included 49 localities in Brazil, French Guiana and Peru and we sampled *c*. 12 individuals per locality to sample variation within and among species and populations. Complementarily, for a posteriori analysis, we included in our dataset information (e.g. geographical coordinates, habitat type etc.) for most of the samples of herbaria collections previously compiled by Vicentini (2007, 2016) and updated here. Herbarium collections included samples from Brazil, Colombia, Guyana, Suriname and Venezuela, and some of these were included in the phylogenetic analyses (Fig. 1). We used samples of *Pagamea acrensis* Steyerm., *P. coriacea* Benth., *P.* m. *igapoana* and *P. duckei* Standl. as the outgroup in the phylogenetic analysis (*P.* m. *igapoana* corresponds to a morphotype of *P. coriacea* in Vicentini, 2016).

Gene selection and probe design

For gene selection and probe design, we downloaded the *Coffea canephora* Pierre ex A.Froehner CDS file from the Coffee Genome hub (http://coffee-genome.org; accessed 1 March 2015). We generated a file containing the genomic locations of the exon boundaries in the genome coding DNA sequences file (CDS file) and genes were split into component exons. The transcriptomes of *Psychotria douarrei* (Beauvis.) Däniker, *Psychotria marginata* Sw. and *Morinda citrifolia* L. were downloaded from the 1kp genome database (http://www.onekp.com/samples/list.php; accessed 1 March 2015), converted to Illumina 1.8+ quality scores (Sass *et al.*, 2016) and cleaned to remove adapters, contaminants, low-complexity sequences and PCR duplicates (Singhal, 2013). The transcriptomes were used to design the bait sequences based on the *C. canephora* genome and annotated exon boundaries as in Sass *et al.* (2016). Exons were filtered to remove those that showed low to zero sequence divergence between the three transcriptomes, thereby including only exons with SNPs in the hope of increasing the likelihood that these markers would have phylogenetic utility at the species-complex level. Briefly, the cleaned transcriptome reads were aligned to the *C. canephora* exons using the software NovoAlign (NovoCraft: http://.novocraft.com; accessed 1 March 2015) setting –t -502 to allow divergent sequences to map; single nucleotide polymorphisms (SNP) were called using SAMtools v.0.1.18 (Li *et al.*, 2009) and VarScan 2.3.6 (Koboldt *et al.*, 2012), and a consensus sequence was generated. The alignment process was repeated a second time with –t 200 using the new consensus sequences as reference. Probes were identified if any of the three transcriptomes had read coverage over at least 150 base pairs of a *C. canephora* exon. To increase the likelihood of generating probes with phylogenetically useful information, when multiple sequences were available per exon, individuals were compared by BLAST and exons with fewer than four nucleotide changes between individuals were eliminated. The total gene list was filtered as in Sass *et al.* (2016), but an additional filter was applied to remove any exons with BLAST hits to any published full plastid or mitochondrial genomes available. Four hundred and fifty-one loci from 341 genes were identified for use as probes. Additionally, we also included the internal transcribed spacer (ITS) and two plast regions (*rps16* and *rpl20-rps12*) because these regions were previously used in the phylogenetic analyses of *Pagamea* (Vicentini, 2016). Finally, the probes were printed out as 60mer oligos at a 1× tiling density on each of three Agilent 1M microarray chips (Sass *et al.*, 2016; Agilent part G3358A). All analyses were conducted in the supercomputer of the Berkeley Research Computing (BRC) at the University of California, Berkeley.

DNA extraction and library preparation

We extracted DNA from dry tissues of 179 samples following a modified 6% CTAB protocol described in Vicentini (2007). For library preparation, we obtained between 1.0 and 1.5 μg of DNA per sample. For each sample, the DNA was sonicated using a Bioruptor® (Diagenode, Liège, Belgium), resulting in fragments of 300 base pairs on average (100–500 bp). DNA fragment sizes were visualized on 1.6% agarose gels. The subsequent steps of blunt-end repair, adaptor ligation, adapter fill-in and indexing PCR were done according to Meyer & Kircher (2010), except that dual indexes were added as described in Kircher, Sawyer & Meyer (2012). DNA enrichment was performed in two (or more) separate PCRs with as few cycles as possible (between six and ten) to limit PCR bias. The PCR products for the same sample were pooled and measured by Nanodrop®, with final concentrations of 20–50 ng/μl in 40 μl per sample.

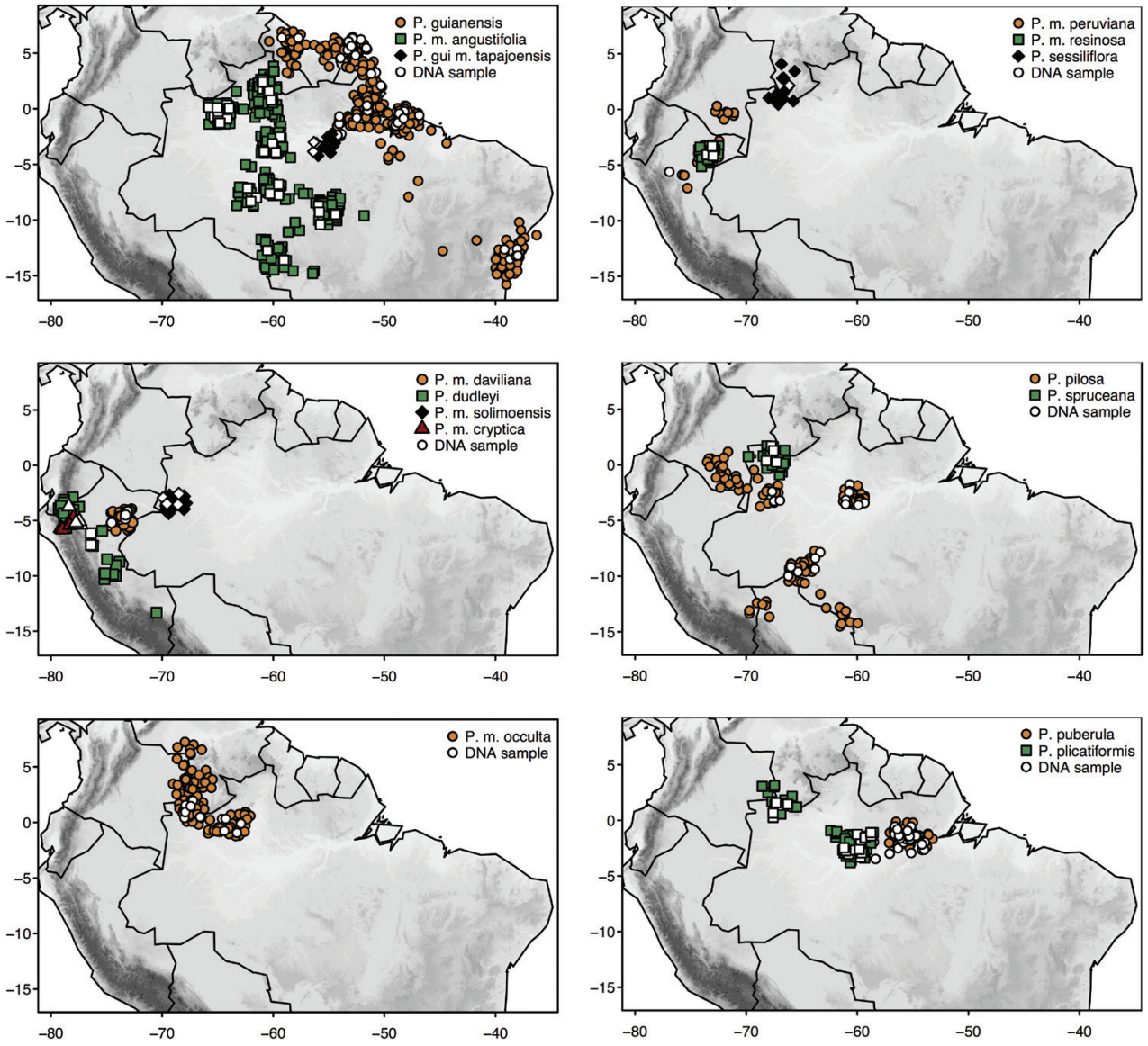

Figure 1. Map of the central-north region of South America showing all known collection points for the PGC. White symbols refers to samples used in the phylogenetic analysis. P. gui m. tapajoensis refers to *P. guianensis* m. tapajoensis.

A total of 433.33 ng of DNA per sample was obtained for the subsequent step. To perform DNA hybridization, three library pools of 60 samples each were created (one per microarray chip), each pool with 26 μg DNA (60 samples × 433.33 ng/sample). After hybridization, each captured library was enriched by performing a limited amount of PCR amplification. Final concentrations were measured by Qubit® and the length distribution of DNA sequences was checked using a Bioanalyzer®. We analysed average length and concentration parameters for each captured library and after pooling all captures, for both pre- and post-hybridization products. Finally, to test the success of the exon-capture experiment, we conducted a qPCR using primers for three target and three non-target genes to be amplified in a reaction with the captured library, and in a reaction with the control DNA (i.e. the original library before hybridization). After confirming the success of DNA capture, all libraries were pooled and a total of 1.8 μl of 4.14 ng/μl averaging 321 basepairs was sequenced in an Illumina® HiSeq® 4000 platform, at the Vincent J. Coates Genomics Sequencing Facility at the University of California, Berkeley.

Pre-process captured reads and sequences alignment

After sequencing, raw sequences were cleaned by removing low quality sequences, adaptors with

indexing barcodes, contaminants and PCR duplicates (Singhal, 2013). Individual references were created by an iterative SNP calling process to the baits previously generated modified slightly from Sass *et al.* (2016). SNP calls were made and genotypes were called only in areas with greater than 5× coverage (for more details, commands and scripts see *github*/*chodon*/*Pagamea* and Sass *et al.*, 2016). Finally, we selected each gene region from each sequenced sample to create one file per gene, with all sequenced samples. With sequences in hand, we created an interactive script to run the software MAFFT for all genes, individually, using the G-INS-1 method, which is a slow but accurate algorithm for sequence alignment (Katoh & Standley, 2013).

Phylogenetic analyses and species tree

We conducted maximum likelihood and Bayesian phylogenetic analyses based on concatenation methods using RaxML (Stamatakis, 2014) and ExaBayes software (Aberer, Kobert & Stamatakis, 2014), for a set of 431 loci from 174 individuals. Given that phylogenetic trees inferred from concatenated regions do not take into account the possible differences between the phylogenetic history of each region, we also estimated these relationships using a species tree method in Astral-II (Mirarab & Warnow, 2015). Our phylogenetic approaches aimed to test for the monophyly of the complex and the previously recognized taxa in the complex and to clarify the phylogenetic placement of populations/samples for which we had no clear previous phylogenetic hypothesis.

We ran maximum likelihood analyses using a GTRCAT model with 100 bootstrap replicates in the software RAxML-HPC v.8 available at the Cypress server (https://www.phylo.org; accessed 10 April 2016). For the ExaBayes analysis, we ran two Metropolis-coupling replicates with four coupled-chains (each with three heated chains) for 1×10^6 MCMC generations, sampled every 500 generations. The following parameters were set according to default configurations: uniform prior for the topology; exponential branch length prior, an exponential prior with parameter λ; reversible matrix with a Dirichlet prior as rates of substitution in the GTR matrix ($N = 6$); uniform rate heterogeneity, where α values have uniform prior probability in the range [0, 100], and state frequencies with Dirichlet priors for state frequencies in a GTR matrix ($N = 4$). All parameters except branch lengths were unlinked across all partitions. The posterior distribution of the parameters were checked for convergence after the average standard deviation of split frequencies (ASDSF) dropped to <5%. Estimated sample size (ESS) for all parameters and branch lengths were summarized with the Exabayes 'postProcParam' tool and confirmed as sufficiently sampled after analysing in Tracer v.1.6 (Aberer *et al.*, 2014): ESS > 200 for most of them, and between 100 and 200 for some of them. Finally, we generated a majority-rule consensus tree with the Exabayes 'consense' tool after a 25% burnin.

We used the software ASTRAL-II v4.10.2 (Mirarab & Warnow, 2015) to infer a species tree from a set of 174 individuals and 429 loci (we excluded the two plastid gene regions due to haploidy). This software was developed to perform coalescent-based analyses over large datasets and has been shown to be statistically robust under the multi-species coalescent model, especially in situations in which incomplete lineage sorting is high (Chou *et al.*, 2015; Mirarab & Warnow, 2015). These methods give a high probability that a true topology will be recovered given a large enough number of true gene trees (Mirarab & Warnow, 2015). We inferred a species tree from 429 individual unrooted gene trees (one tree per loci), previously estimated with 100 bootstrap replicates in RaxML v8 (Stamatakis, 2014). Samples used for phylogenetic and species tree analysis are listed in Table S1 (Supporting Information).

Coalescent species delimitation

We used BPP v3.2 (Yang & Rannala, 2010, 2014) to test species limits hypotheses among the 14 *Pagamea* clades of the PGC detected in our previous phylogenetic analyses, i.e. all nominal species (putative species). This most recent version of the program can run the reversible-jump MCMC species delimitation algorithm and simultaneously estimate a species tree. Briefly, a subtree pruning and regrafting algorithm is used to vary the species tree topology, and species hypotheses are tested by collapsing the branches in these topologies and comparing their posterior probabilities under a multi-species coalescent model (Rannala & Yang, 2015). Since BPP tests all possible species tree topologies, it also tests the hypotheses that any combination of two or more proposed species could actually belong to one single species. Thus, unlike previous BPP versions, the fact that the species tree topology can change eliminates the concern of over-estimating lineages limits (Leaché & Fujita, 2010; Caviedes-Solis *et al.*, 2015).

After phylogenetic analysis, we updated our initial 11 species hypothesis to 14 putative species to be tested in BPP: *P. dudleyi*, *P. guianensis*, *P.* m. *cryptica*, *P.* m. *peruviana*, *P.* m. *resinosa*, *P. spruceana*, *P.* m. *occulta*, *P. plicatiformis*, *P. pilosa*, *P. puberula*, *P. sessiliflora* and the new detected clades in the phylogenetic analysis *P.* m. *angustifolia*, *P.* m. *daviliana* and *P.* m. *solimoensis*. After several trials using different parameters, we used a gamma prior of ~G (14, 1000) for population size (theta, θs), and

~G (2, 1000) for the age of the root in the species tree (tau, τ0), and the Dirichlet prior (Yang & Rannala, 2010: Equation 2) for other divergence time parameters. The gamma prior G (α, β) has mean α/β, so the theta prior G (14, 1000) corresponds to 14 differences per kilo base (0.014), whereas the tau prior ~G (2, 1000) corresponds to 0.2% sequence divergence. In other words, we used priors that would correspond to a relatively large population size and relatively shallow divergence times. Based on information from the literature, those priors reflect a biologically meaningful scenario for the groups: diversification events took place in the last 18 My in *Pagamea* and in the last 4 My in the PGC (Bremer & Eriksson, 2009; Vicentini, 2016). Moreover, and probably more important in this context, they are empirically consistent, since convergence evaluation among different BPP runs is based on the similarity of the results from different runs. Our final prior choice returned the same results in almost every run, whereas other priors delivered different results on different runs.

For the BPP species delimitation analysis we used the 426 nuclear loci, excluding two plastid regions and three nuclear gene regions with too many missing data and also excluding the outgroup. Because of computational constraints, we used two or three individuals per species sampled from different populations, representing each clade recovered in the phylogenetic analyses of all sampled individuals. This sampling strategy is in accordance with the species delimitation model implemented in BPP, in which even one individual is sufficient for robust species delimitation given that enough loci (>50) are sampled (Yang & Rannala, 2010; Zhang *et al.*, 2011). We ran the analysis for 1×10^6 MCMC generations, taking samples every five generations and using 1×10^5 burn-in generations. To conform to the multi-species coalescent model and avoid unwarranted influences in the likelihood calculations, we ran the analysis excluding gaps and ambiguous sites from the alignment (cleandata = 1), using both available reversible-jump MCMC species delimitation algorithms (algorithms 0 and 1; Yang & Rannala, 2010). Finally, to evaluate convergence, for each analysis we conducted at least three independent runs starting with random tree models.

Morphological analyses

We analysed a dataset containing only vegetative characters (no reproductive characters) because most of our specimens were collected sterile. We excluded from the analysis all species with few samples (*P. dudleyi*, *P.* m. *cryptica*, *P.* m. *resinosa*, *P.* m. *peruviana* and *P. sessiliflora*). The vegetative characters analysed were 'leaf area', 'number of secondary veins' and 'leaf-shape PCA axis' representing leaf contour shapes generated with the software SHAPE v.1.3 (Iwata & Ukai, 2002). To obtain leaf images as input files for SHAPE, we scanned all samples (three leaves per sample) in a high-resolution scanner and images were converted to a black and white object of known size (based on the area of the image). SHAPE can delineate any type of shape with a closed two-dimensional contour based on the elliptic Fourier descriptors (EFD; Kuhl & Giardina, 1982). The program extracts the contour shape from leaf images, delineates the contour shape with the EFDs and performs a principal component analysis (PCA) of the EFDs for summarizing the shape information (Iwata & Ukai, 2002). Leaf area was also calculated by SHAPE and we counted the number of secondary veins from leaf images.

We ran a multidimensional scaling analysis (MDS) on a Euclidean distance morphological matrix to evaluate whether species grouped in morphological space, using the functions 'cmdscale' and 'dist' from the package 'stats'. To test if species could be discriminated based on their morphology, we applied classificatory analysis using both support vector machines (SVM) (Cortes & Vapnik, 1995) and naiveBayes (NB) classifiers with functions 'svm' and 'naiveBayes' from package 'e1071' (Meyer *et al.*, 2017). We ran analyses for the whole dataset and for some clades from the species tree generated from Astral-II: (1) *P.* m. *daviliana* +*P.* m. *solimoensis*; (2) *P.* m. *occulta* + *P. pilosa* + *P. spruceana* and (3) *P.* m. *angustifolia* + *P. guianensis* + *P. guianensis* m. *tapajoensis* + *P. plicatiformis* + *P. puberula*. These tests were performed to assess if the power of prediction increases in the classifier models when species are compared only with their close relatives. We conducted all analyses in R (R Core Team, 2017).

Near-infrared spectral signature

Near-infrared (NIR) leaf spectroscopy data have been shown to be useful in separating closely related species of Lecythidaceae (Durgante *et al.*, 2013) and Burseraceae (Lang *et al.*, 2015) in the Amazon. Here, we applied NIR leaf spectroscopy to test whether species of the PGC can be discriminated and if the discriminant model agrees with our putative species inferred from our Bayesian species delimitation test. The NIR dataset included 6888 spectra from 575 individuals from 49 sampling sites (i.e. populations) of nine species in the PGC. We excluded from the analysis all species with few samples (*P. dudleyi*, *P.* m. *cryptica*, *P.* m. *resinosa*, *P.* m. *peruviana* and *P. sessiliflora*). For each individual 12 spectra were collected from three leaves (two adaxial and two abaxial for each leaf) using a spectroscopy analyser from Thermo Fisher Scientific, model Antaris II. Each spectrum consists of 1557 values of near-infrared absorbance sampled at intervals of 4 cm^{-1} for

wavelengths 4000–10 000 nm from dried leaves. Linear discriminant analysis (LDA) was used to test whether a specimen of a given species could be identified by NIR spectral data. We performed two different cross-validations (leave one out): one in which the identity of a specimen was predicted based on an LDA model that included other individuals of the same population (sampling ranging from eight to 17 individuals per population) and a second test in which individuals of the same population were absent in the LDA model. All individuals from all populations for each species were tested. In the NIR dataset, individuals are represented by multiple spectra, which were collected to minimize possible effects of outlier spectra on the prediction models. Intra-individual variation may be caused by the reading position on the leaves, leaf age, shading conditions, fungi, epiphylls, reading errors etc. The use of mean spectra per individual (Durgante *et al.*, 2013; Lang *et al.*, 2015) may reduce, but will not eliminate the effects of outlier spectra. Here, we conducted LDA analysis with all spectra per individual. The species prediction and the posterior probability of each spectrum were annotated, and an individual was considered correctly identified if the majority of the spectra were correctly identified, considering only spectra with Posterior Probability > 0.95. These analyses were implemented in R (R Core Team, 2017).

Ecological niche modelling, niche overlap and habitat characterization

We conducted ecological niche modelling (ENM) to estimate habitat suitability for species based on their ecological niches. We downloaded 19 climatic variables from 'WordClim Global Climate Data' (*worldclim.org/version2*; accessed 23 November 2017) and 16 from 'Environmental Rasters from Ecological Modeling' (*envirem.github.io/*; accessed 23 November 2017), two soil variables (arenosols and podzols, because our species are restricted to sand soils) from 'World Soil Information' (ftp://ftp.soilgrids.org/) and two topographic variables (elevation and slope) from 'Shuttle Radar Topography Mission' (*dds.cr.usgs.gov/srtm/*; accessed 10 June 2016) to be used as predictor variables in our models. We calculated pairwise correlation between all environmental predictors and excluded the variables with pairwise correlation > 0.7, using the function 'vifcor' from package 'usdm'. A subset of 12 retained variables was used as predictors in the models. We cropped all rasters using an area extent defined by adding 5° to the minimum and maximum latitude and longitude coordinate values for each sample of the PGC. We modelled the distribution of all species using the function 'ENMevaluate' of the package 'ENMeval', implemented to execute automated runs and evaluations of the ecological niche models (Muscarella *et al.*, 2014) and the function maxent (version 3.3.3k; Phillips & Dudík, 2008) with the package 'dismo' (Hijmans *et al.*, 2014). We randomly selected 10 000 points from environmental variables to be used as background. We set the function to allow for 'linear', 'quadratic' and 'product' features and used the method 'random-*kfolds*' for partitioning occurrence data into five-fold cross-validation. We selected the model with the higher AUC to then run maxent with the same arguments of the selected model and calculated the cumulative probabilities to generate the maps of relative habitat suitability for each species.

To test for niche equivalency/similarity between closely related species, we applied the approach proposed by Broennimann *et al.* (2012) using the function 'ecospat.niche.equivalency.test' from the package 'ecospat'. This function quantifies niche overlap between species using the statistics D and I, which varies between 0 (no overlap) and 1 (complete overlap) (Warren, Glor & Turelli, 2008). For each pair of species, we first applied a PCA (using the same environmental variables as in the ENM analysis) to calculate the scores for the species in a two axis ordination, using 2000 points randomly selected as background. Then, for each species we created an occurrence density grid with the function 'ecospat.grid.clim.dyn'. Finally, we calculated the niche overlap and the significance of the results by performing 1000 replications. If the observed niche overlap was less equivalent/similar than expected by chance, we accepted the hypothesis of niche divergence.

We characterized species habitat in relation to vegetation structure and soil flooding. Vegetation structure was described as savanna-like (< 8 m), low forest (8–12 m) or tall forest (> 12 m), and flooding was described as non-flooded, moist (usually by groundwater), flooded (< 6 months) and long-term flooded (> 6 months), adapted from Vicentini (2016).

RESULTS

High-throughput exon-capture sequencing

Our high-throughput exon-capture experiment resulted in all loci sequenced from 454 baits used (451 designed plus ITS and the two plastid regions). After removing loci with too many missing data, 431 loci from 321 genes (from the 12 chromosomes of the *C. canephora* reference genome) were concatenated in a matrix of 174 individuals and 196 925 base pairs, totaling 34 265 124 characters. The average loci length was 455 base pairs (range: 101 to 1534), the average number of potentially informative sites per marker was 69, and the total number of ambiguities was low (0.04%).

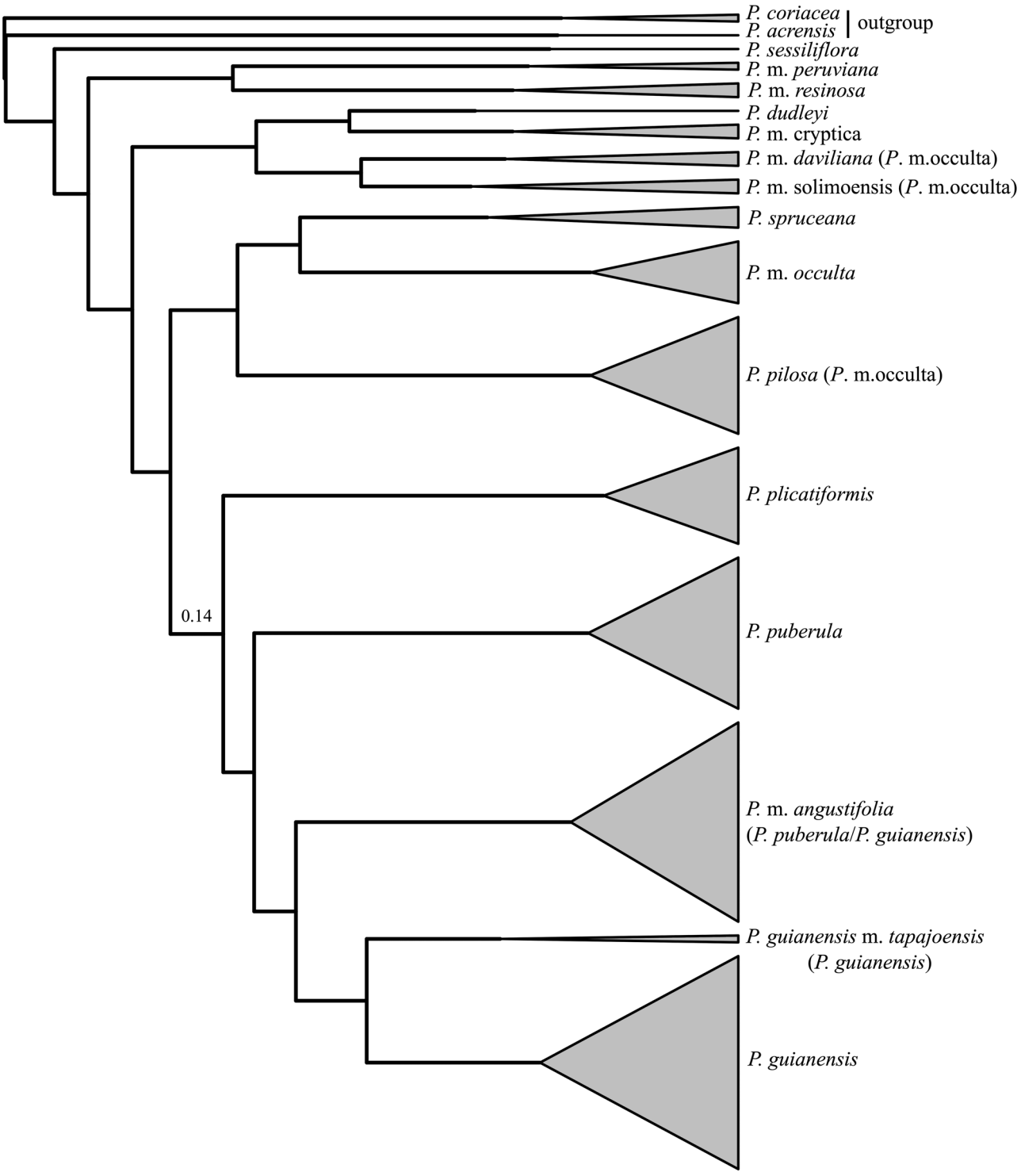

Figure 2. Species tree of the PGC generated with the multi-species coalescent model in the software Astral-II. Individuals from the same taxon were grouped into triangles with sizes proportional to the number of tips in the tree. Nodes with Bootstrap > 95% are omitted. *Pagamea coriacea* and *P. acrensis* are outgroups. Names inside parenthesis refer to the circumscription proposed by other authors in their latest reviews (see Supporting Information, Fig. S5).

Are the *Pagamea guianensis* complex (PGC) and currently described species monophyletic?

The PGC was recovered as monophyletic with strong support in RaxML and Astral-II analyses (maximum likelihood: ML = 0.96 and bootstrap support: BS = 100, respectively), but with lower support in Exabayes analysis (posterior probability: PP = 0.65) (Supporting Information, Figs S1–S3). Phylogenetic analyses on RaxML and Exabayes (based on concatenated data) produced almost identical topologies, with minor differences in internal relationships in species clades. Previously described species *P. dudleyi*, *P. pilosa* and *P. plicatiformis* were monophyletic in all analyses (Fig. 2, Supporting Information Figs S1–S3). On the other hand, *P. puberula* was paraphyletic in relation to *P. guianensis* in all phylogenetic trees, because some populations from the central Amazon were more closely related to *P. guianensis* than to *P. puberula*. We informally named the clade that includes these populations and others with the similar morphology as '*P.* m. *angustifolia*'. *Pagamea guianensis* is reciprocally monophyletic to '*P.* m. *angustifolia*' in all phylogenetic trees and contains two subclades, one corresponding to populations from the Guiana and Brazilian Shields and the Atlantic Coast, and one corresponding to a morphologically distinct small population isolated from the others by the Amazonas and Tapajós Rivers, here informally named *P. guianensis* m. *tapajoensis*. Previously proposed (but not described; Vicentini, 2016) species *Pagamea* m. *macrocarpa* (here *P.* m. *cryptica*), *P.* m. *peruviana* and *P.* m. *resinosa* were monophyletic in all trees, whereas *P.* m. *occulta* was polyphyletic in all phylogenetic trees, resulting in three newly detected independent cryptic lineages, here informally named *P.* m. *occulta*, *P.* m. *daviliana* and *P.* m. *solimoensis* (Table 1). Species tree analysis based on the multi-species coalescent model (on Astral-II) generated a highly resolved species tree in which previously and newly recognized taxa corresponded to clades with strong support (PP > 0.95) (Fig. 2).

How many species are supported by the coalescent species delimitation analyses?

After updating our species hypotheses, we tested whether the new putative species were recovered by the coalescent species delimitation analysis in BPP. This analysis recovered the 14 hypothesized species with high posterior probability (PP = 1), although in

Table 1. General framework showing the phylogenetic status of species as currently circumscribed and the new phylogenetic status after updating species hypothesis based on the results generated in this work for the *Pagamea guianensis* complex. **Pagamea* m. *occulta* and *P. spruceana* were recovered as one or two lineages depending on the BPP run (see in the results)

Currently proposed circumscription in (Vicentini, 2016)	Phylogenetic status in Vicentini (2016)	Phylogenetic status in the new trees	New proposed Circumscription	New phylogenetic status	Sympatric with sister species or clade?
P. dudleyi	monophyletic	monophyletic	*P. dudleyi*	monophyletic	sympatric
P. guianensis	paraphyletic/polyphyletic	paraphyletic to *P. puberula*	*P. guianensis*	monophyletic	allopatric
			P. guianensis m. *tapajoensis*	monophyletic	allopatric
	paraphyletic		*P.* m. *angustifolia*	monophyletic	sympatric
P. m. *macrocarpa*	monophyletic	monophyletic	*P.* m. *cryptica*	monophyletic	sympatric
P. m. *occulta* (*P. pilosa*)	polyphyletic	polyphyletic	*P. pilosa*	monophyletic	allopatric
			P. m. *occulta*	monophyletic*	sympatric
			P. m. *daviliana*	monophyletic	allopatric
			P. m. *solimoensis*	monophyletic	allopatric
P. m. *peruviana*	monophyletic	monophyletic	*P.* m. *peruviana*	monophyletic	sympatric
P. m. *resinosa*	monophyletic	monophyletic	*P.* m. *resinosa*	monophyletic	sympatric
P. m. *spruceana*	paraphyletic	monophyletic	*P. spruceana*	monophyletic*	sympatric
P. plicatiformis	paraphyletic	monophyletic	*P. plicatiformis*	monophyletic	sympatric
P. puberula		paraphyletic to *P. guianensis*	*P. puberula*	monophyletic	sympatric
P. sessiliflora	single accession	monophyletic	*P. sessiliflora*	monophyletic	sympatric

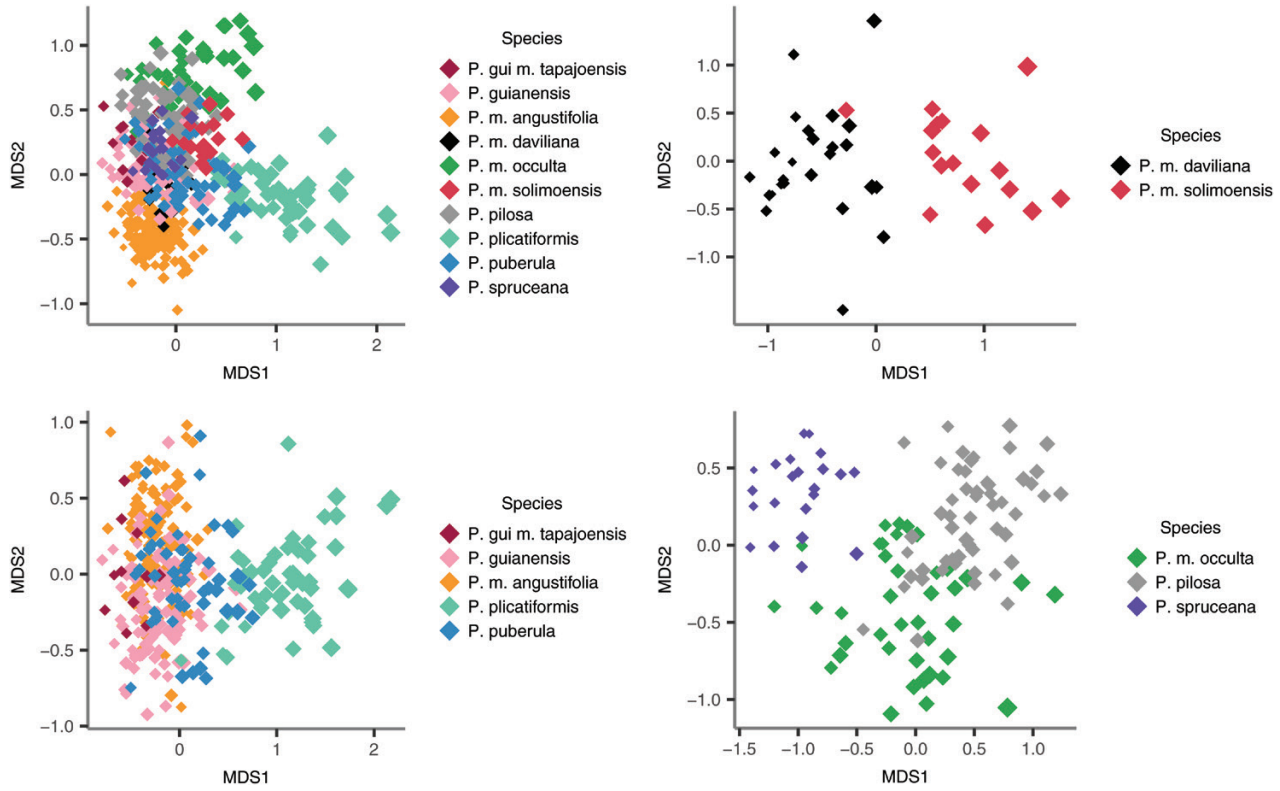

Figure 3. Multidimensional Scaling analysis (MDS) of morphological data represented by the variables 'leaf area', 'number of secondary veins' and 'leaf-shape PCA axis' for: (**A**) all species of the PGC analyzed together; (**B**) clade *Pagamea* m. *daviliana* + *P.* m. *solimoensis*; (**C**) clade *P.* m. *angustifolia* + *P. guianensis* + *P. guianensis* m. *tapajoensis* + *P. plicatiformis* + *P. puberula*; (D) clade *P.* m. *occulta* + *P. pilosa* + *P. spruceana*.

two out of six runs only 13 species were delimited (PP = 1) because of the high posterior probability for the combination *P. spruceana* + *P.* m. *occulta* as a unique species.

Are species morphologically different?

The MDS analysis of leaf shape and size showed morphological overlap among most species (Fig. 3A). However, we found strong patterns of species clustering in the morphological space when clades were analysed separately (Fig. 3B–D). Discriminant models correctly predicted 85 and 79% (SVM and NB models, respectively) of the samples into their respective species when analysing all species together (Table 2, Fig. 4). Although, on average, the SVM model performed better, the NB model yielded a considerably high value of species assignment for *P.* m. *daviliana* (the only taxon with low prediction in the SVM analysis) when compared to the SVM model (0.84 and 0.48%, respectively). We found even higher values of species assignment when applying the SVM model to each clade separately: 100 and 94% for Clade 1

Table 2. Species classification analysis applied to the morphological variables 'leaf area', 'number of secondary veins' and 'leaf-shape PCA axis' in the *Pagamea guianensis* complex, using support vector machine (SVM) and naiveBayes classifiers. SVM analysis (marked with *) was also applied separately for the clades (1) *P.* m. *daviliana* + *P.* m. *solimoensis*; (2) *P.* m. *angustifolia* + *P. guianensis* + *P. guianensis* m. *tapajoensis* + *P. plicatiformis* + *P. puberula* and (3) *P.* m. *occulta* + *P. pilosa* + *P. spruceana*

	Taxon	SVM*	SVM	NB
Clade 1	*P.* m. *daviliana*	1.00	0.48	0.84
	P. m. *solimoensis*	0.94	0.82	0.88
Clade 2	*P.* m. *occulta*	0.95	0.90	0.85
	P. pilosa	0.92	0.77	0.62
	P. spruceana	1.00	1.00	0.95
Clade 3	*P. guianensis*	0.86	0.79	0.62
	P. guianensis m. *tapajoensis*	0.55	0.54	0.82
	P. m. *angustifolia*	0.99	0.98	0.90
	P. plicatiformis	0.98	0.96	0.90
	P. puberula	0.76	0.69	0.56
Total		0.95	0.85	0.79

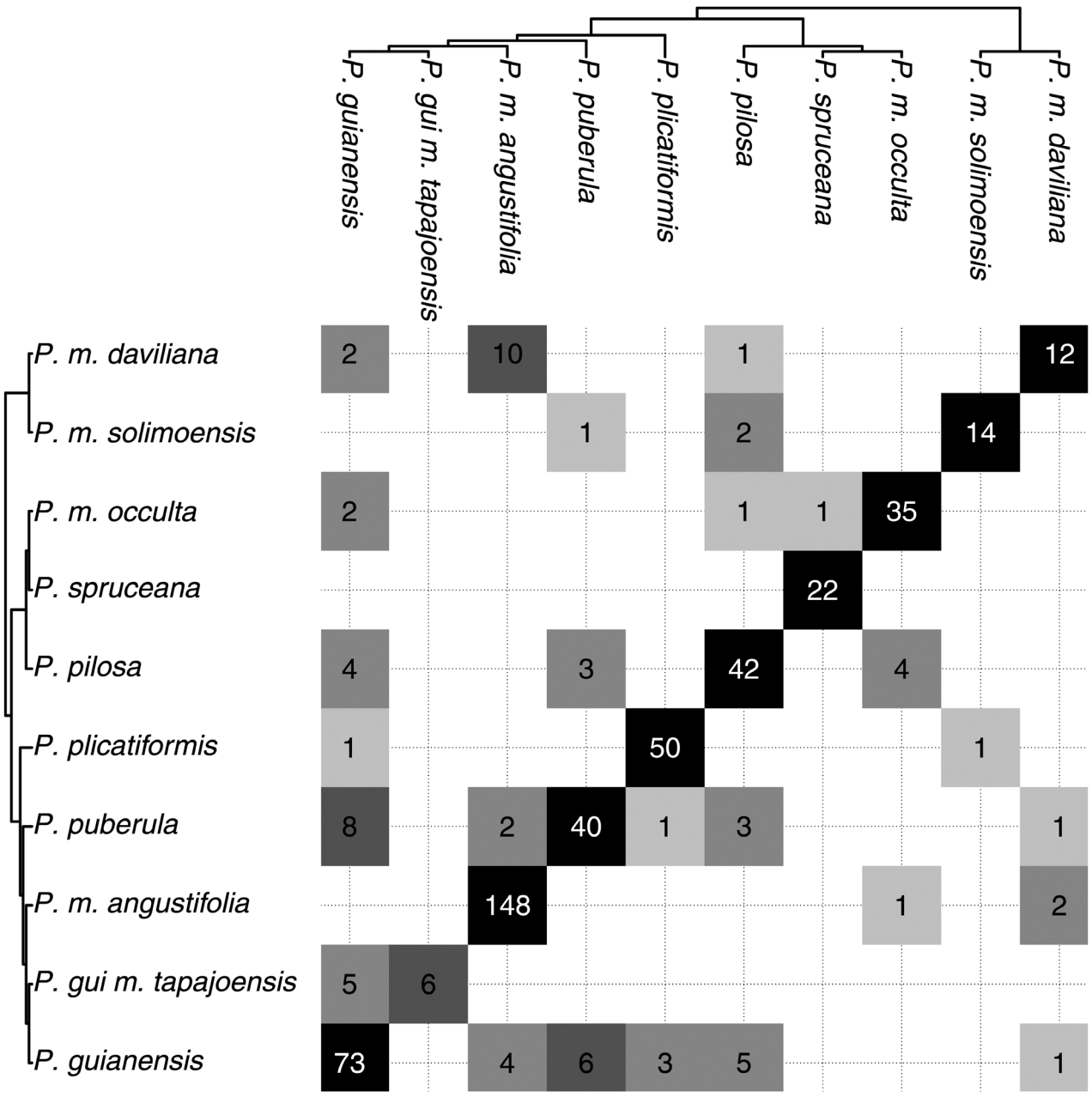

Figure 4. Matrix of confusion resulted from the species classification analysis applied to the morphological variables 'leaf area', 'number of secondary veins' and 'leaf-shape PCA axis', using a Support Vector Machine (SVM) classifier. P. gui m. tapajoensis refers to *P. guianensis* m. *tapajoensis*.

(*P.* m. *daviliana* + *P.* m. *solimoensis*); 95, 92 and 100% for Clade 2 (*P.* m. *occulta* + *P. pilosa* + *P. spruceana*); 99, 86, 55, 98 and 76% for Clade 3 (*P.* m. *angustifolia* + *P. guianensis* + *P. guianensis* m. *tapajoensis*, *P. plicatiformis* + *P. puberula*) and 0.96, 0.97, 0.92, 0.94 and 100% for both Clades 1 and 2 analysed together (*P.* m. *daviliana*, *P.* m. *solimoensis*, *P.* m. *occulta*, *P.* pilosa and *P. spruceana*, respectively) (Table 2).

Are species spectrally different?

Linear discriminant analysis of the NIR spectral data for leaves performed well for both models. In the first, less conservative model, in which other individuals of the same population are present in the model, 99.8% of the samples were correctly identified (Fig. 5). In the second model, in which samples of the same population were not included in the LDA model, the number of correct predictions dropped to 97%.

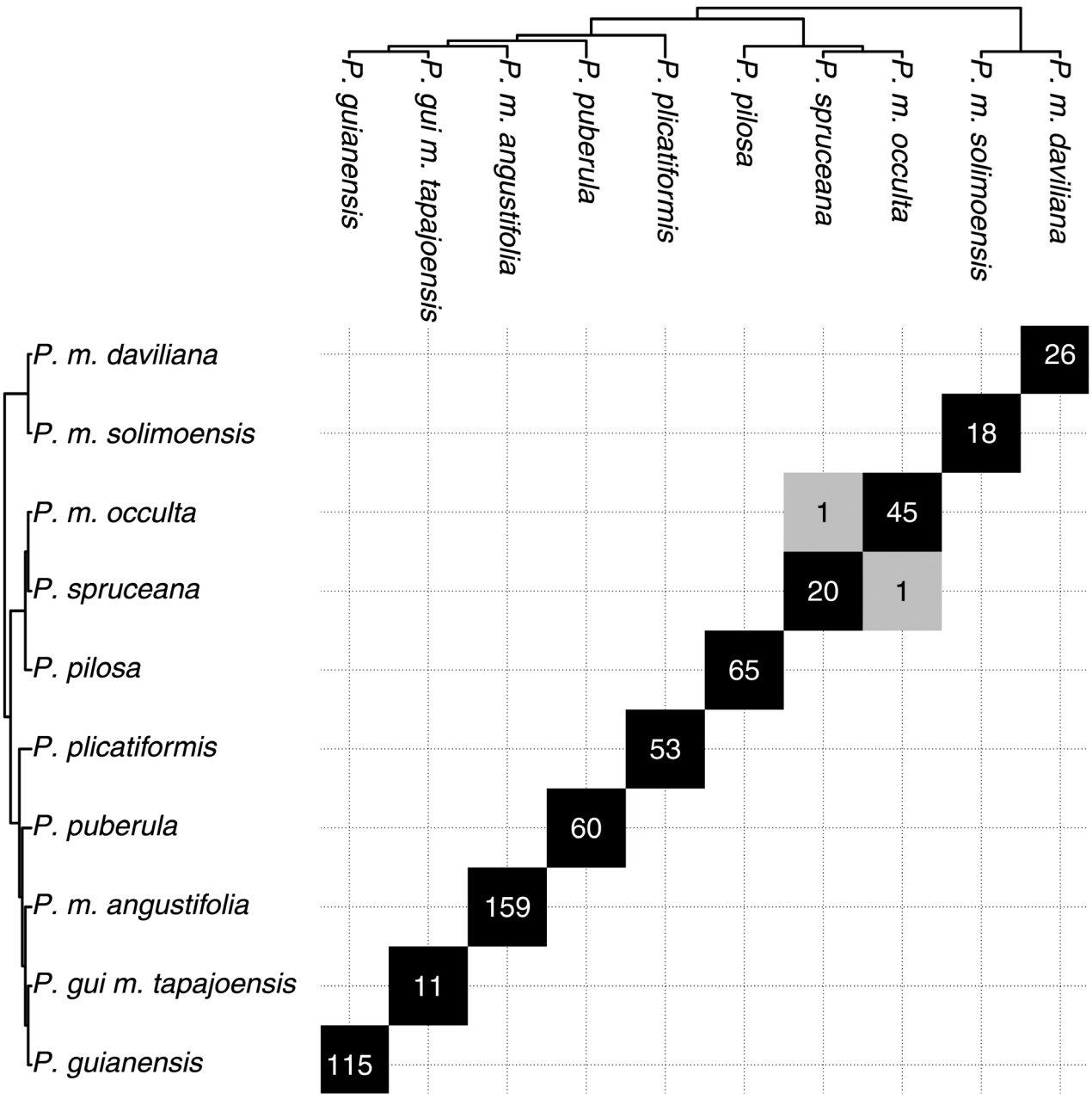

Figure 5. Matrix of confusion resulted from the species classification analysis applied to the NIR spectral data using a Linear Discriminant Analysis (LDA). P. gui m. tapajoensis refers to *P. guianensis* m. *tapajoensis*.

ARE SPECIES ECOLOGICALLY DIFFERENT?

The main variables explaining relative habitat suitability for the species were 'precipitation of coldest quarter' (for *P.* m. *angustifolia*, *P.* m. *occulta*, *P. plicatiformis*, *P. puberula*), 'precipitation seasonality' (for *P.* m. *cryptica*, *P.* m. *daviliana*, *P.* m. *peruviana*, *P.* m. *resinosa* and *P. sessiliflora*), 'podzols' (for *P.* m. *solimoensis* and *P. spruceana*), 'slope' (for *P. dudleyi*), 'precipitation of warmest quarter' (for *P. pilosa*) and 'count of the number of months with mean temperature > 10 °C' (for *P. guianensis*) (Table 3; Supporting Information, Fig. S4). Values of niche overlap calculated between species were in general low (< 0.5) and less equivalent/similar than random ($P < 0.05$; Fig. 6), thus indicating niche divergence. The only exceptions were for the pairs *P. dudleyi*/*P.* m. *cryptica* (sympatric in the sub-Andean region) and *P. plicatiformis*/*P. sessiliflora* (sympatric in the Upper Rio Negro region), where the null hypotheses of niche equivalency/similarity were not rejected ($P > 0.05$).

Table 3. Contribution of each variable in the ecological niche modelling (ENM) per species generated in Maxent. Columns 1 to 12 refer to ecological variables according to their original definition (1–3 from Envirem, 4 from SRTM, 5–6 from Soilgids, 7–12 from Bioclim2). 1. Count of the number of months with mean temp >10.2. Mean monthly PET of driest quarter. 3. Mean monthly PET of wettest quarter. 4. Terrain slope. 5. Arenosols. 6. Podzols. 7. Mean diurnal range (mean of monthly (max temp - min temp)). 8. Isothermality (BIO2/temperature annual range) (×100). 9. Precipitation of wettest month. 10. Precipitation seasonality (coefficient of variation). 11. Precipitation of warmest quarter. 12. Precipitation of coldest quarter. AUC values for the ENM analysis per species. PET = Potential Evapotranspiration.

Taxon	1	2	3	4	5	6	7	8	9	10	11	12	AUC
P. dudleyi	4.8	0.3	0.8	**34.6**	1.1	5.7	18.9	0.4	0.3	20.5	1.2	11.4	0.94
P. guianensis	**39.3**	0.8	1.0	7.4	2.0	4.6	33.2	1.3	0.5	0.4	3.1	6.3	0.81
P. m. *angustifolia*	0.0	0.6	8.5	2.9	17.5	9.1	4.8	7.3	**14.2**	2.1	10.2	22.6	0.93
P. m. *cryptica*	2.5	0.1	0.2	25.1	1.2	10.9	2.4	16.2	8.2	**33.2**	0.1	0.1	0.99
P. m. *daviliana*	1.4	13.8	6.1	1.4	8.9	4.3	0.5	1.8	7.7	**25.8**	10.2	18.2	0.99
P. m. *occulta*	0.0	6.4	2.7	7.4	2.4	7.6	4.8	8.3	0.7	4.1	3.4	**52.2**	0.97
P. m. *peruviana*	0.4	1.6	0.1	2.6	5.3	3.2	4.6	5.2	12.1	**31.8**	18.7	14.5	0.99
P. m. *resinosa*	2.9	1.3	4.6	2.1	13.9	4.3	7.1	1.0	3.8	**37.9**	11.4	9.8	0.99
P. m. *solimoensis*	0.1	12.3	6.8	1.7	3.0	**21.2**	13.0	17.2	2.3	9.3	10.6	2.6	0.95
P. pilosa	0.0	2.1	0.0	8.8	4.9	10.5	10.1	8.5	6.0	1.5	**33.4**	14.1	0.91
P. plicatiformis	0.1	0.6	0.4	5.9	9.9	8.7	0.7	8.1	0.7	8.5	19.5	**36.9**	0.97
P. puberula	0.1	0.1	13.8	6.9	19.8	7.6	2.4	7.9	0.7	1.1	5.1	**34.4**	0.99
P. sessiliflora	1.7	27.0	0.0	2.0	5.8	6.0	1.1	0.8	0.0	**25.9**	10.5	19.3	0.97
P. spruceana	1.6	3.7	0.4	0.7	1.8	**51.9**	0.1	0.2	0.4	15.0	19.3	4.8	0.99

Variation in habitat in relation to vegetation structure and soil water regime was usually pronounced between sister or closely related taxa when in sympatry (Table 4), for example *P.* m. *occulta* and *P. spruceana* in flooded and long-term flooded river banks habitats, respectively, in the upper Rio Negro region; *P. puberula* in dry clay or sand soil in tall forests, *P. plicatiformis* in moist sand soil in low to tall forests and *P.* m. *angustifolia* in open vegetation or shrublands in the Central Amazon, respectively; and *P. dudleyi* in the shrubland and *P.* m. *cryptica* in tall forests in northern Peru, respectively. On the other hand, sympatric taxa *P.* m. *resinosa* and *P.* m. *peruviana* share almost identical habitats in the region of Iquitos, Peru, whereas allopatric sister taxa *P.* m. *angustifolia* and *P. guianensis* share similar habitats in relation to vegetation structure and soil water regime. All proposed new taxa in this study will be published in a separate manuscript with complete morphological descriptions.

DISCUSSION

Our multiple-evidence approach based on genomic, morphological, NIR spectral and ecological data hypothesizes the existence of 15 highly supported lineages in the PGC. These findings increase the number of species in the PGC from seven previously formally described (including *P. spruceana*, recently described in Prata *et al.*, 2016, based on the results here achieved) to 14 species and one subspecies proposed in this work (Supporting Information, Fig. S5). Although many of these species (ten) were recognized in Vicentini (2007, 2016), here we changed circumscriptions for some, accepted and confirmed species hypotheses for others and discovered previously undetected cryptic taxa. Our results highlight the importance of a multiple evidence approach to species delimitation in species complexes for the discovery of cryptic species and for estimating the total number of species in the Amazon. For example, some of the 'hyperdominant' species are potential species complexes (ter Steege *et al.*, 2013), and few cryptic species are treated in the compilation of regional floras discussed by Cardoso *et al.* (2017).

Our sampling design based on the sequencing of several genes from multiple individuals per population and many populations per species enabled the reconstruction of a robust species tree even in the presence of paraphyletic gene trees. For instance, gene trees of *Pagamea* from nuclear (ITS) and plastid (*rps16* and *rpl20-rps12*) markers were conflicting and presented unsolved phylogenetic relationships for some species (especially for those of the PGC; Vicentini, 2016), whereas here we were able to accomodate different gene histories (including the markers used in the *Pagamea* phylogenetic analyses) in strongly supported phylogenetic and species trees. Although the phylogenetic analyses based on different methods recovered similar topologies, the multispecies coalescent delimitation methods (Yang & Rannala, 2010; Mirarab & Warnow, 2015; Rannala, 2015; Yang, 2015) recovered

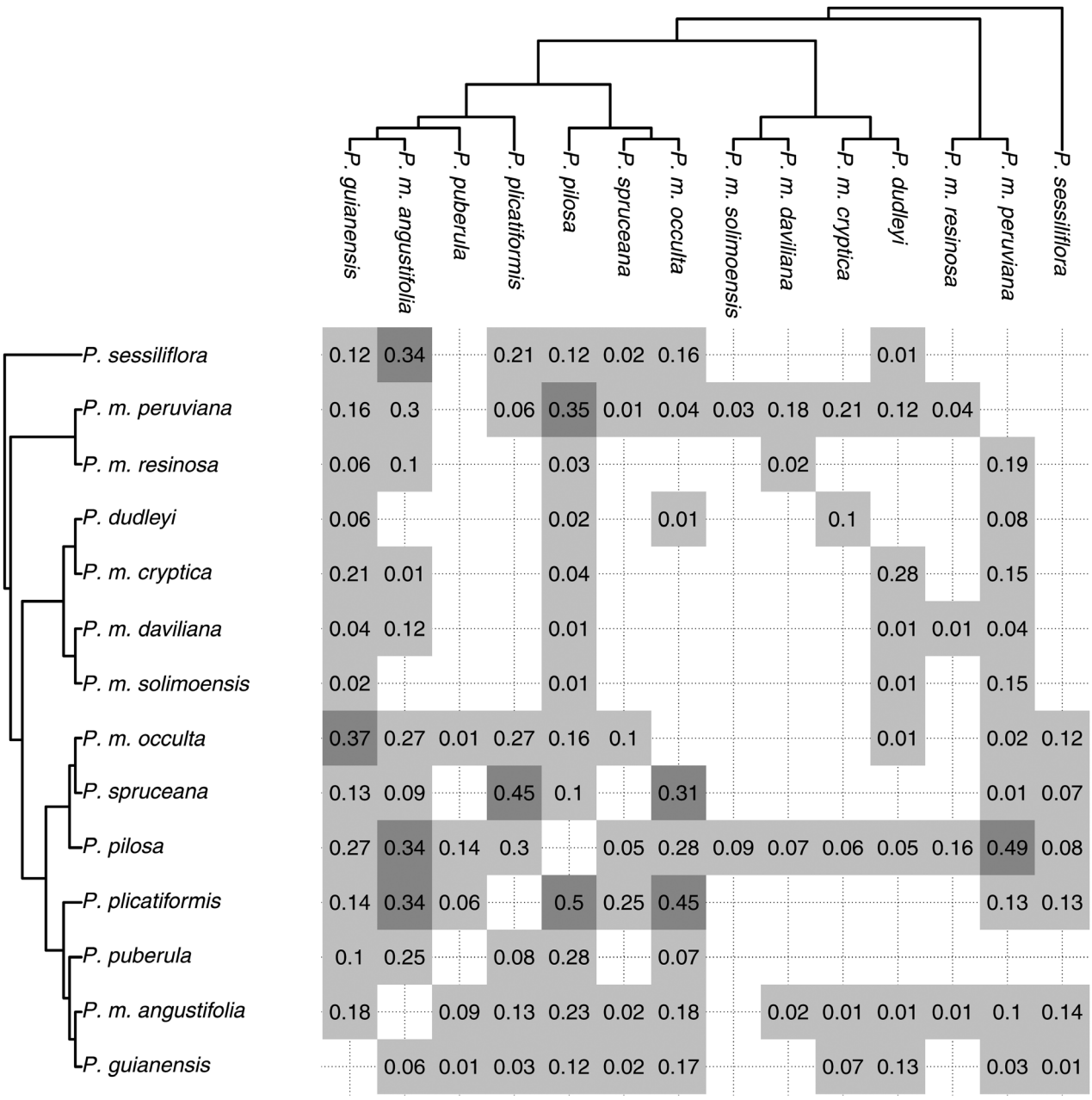

Figure 6. Matrix of confusion with ecological niche overlap values for both I (the lower triangle) and D (the upper triangle) distances. $P < 0.05$ for all values, with exception of *Pagamea dudleyi*/*P.* m. *cryptica* and *P. plicatiformis*/*P. sessiliflora*. Empty cells correspond to values equal to zero.

the tree with the highest values of node support. This may be explained by the fact that multispecies coalescent methods perform better than concatenation methods when the degree of incomplete lineage sorting (ILS) is high and many gene trees conflict (Edwards *et al.*, 2016), which seems to be the case of taxa in the PGC. These results show the importance of the multi-individual and multi-loci sampling and the multispecies coalescent approach when inferring phylogenetic relationships and species limits in species complexes in the extremely diverse and understudied Amazonian flora.

Because the multispecies coalescent delimits population-level structure rather than species boundaries (as in BPP species delimitation test), genomic-based hypotheses about species limits should be validated with multiple sources of evidence (Sukumaran & Knowles, 2017). Thus, after updating and testing our species hypotheses based on genomic

Table 4. Habitat structure (vegetation and flooding regime) of the taxa in the *Pagamea guianensis* complex. *Some collections of *P. puberula* were found in clay soils at the region of Manaus

Taxon	Vegetation structure	Soil
P. dudleyi	shrubland, forest	dry, moist
P. guianensis	open	dry, moist
P. m. *angustifolia*	open	dry
P. m. *cryptica*	tall forest	dry
P. m. *daviliana*	low to tall forest	dry, moist
P. m. *occulta*	open, low to tall forest	flooded
P. m. *peruviana*	tall forest	dry
P. m. *resinosa*	low to tall forest	dry
P. m. *solimoensis*	tall forest	dry
P. pilosa	low to tall forest	moist, flooded
P. plicatiformis	low to tall forest	moist
P. puberula	open, low to tall forest	dry, moist, *
P. sessiliflora	tall forest	dry, moist
P. spruceana	open, low forest	long-term flooded

data, we evaluated whether species were supported by morphological, NIR spectral and ecological data. Here, we assume that it is not a simple matter to propose an objective approach for species delimitation given the nature of the speciation process, in which no order is expected in the appearance of divergences among sister species, such as morphological differentiation, ecological adaptation to new niches and reproductive isolation among other criteria (de Queiroz, 1999, 2007). For instance, it is relatively uncontroversial to accept as distinct species two or more sympatric closely related clades which are morphologically, ecologically and spectrally different, as in the case of *P.* m. *angustifolia*, *P. plicatiformis* and *P. puberula* in Central Amazon and *P.* m. *occulta* and *P. spruceana* in the Upper Rio Negro. On the other hand, the decision about species, subspecies or population-level status is complicated and somewhat subjective when lineages are allopatric and not all criteria are concordant. In this case, we decided to delineate as distinct species when allopatric sister lineages are divergent for all criteria, as *P.* m. *daviliana* and *P.* m. *solimoensis*. Because there is no real limit to the subspecies category (Wilson & Brown, 1953), here we arbitrarily proposed as subspecies a genetically and geographically isolated population of *P. guianensis* that showed some degree of morphological differentiation (captured in the morphological analysis), here named *P. guianensis* m. *tapajoensis*. Finally, in the case of reciprocally monophyletic and geographically isolated clades with no clear morphological difference we interpreted these as genetically structured populations.

Although some species of the PGC overlap morphologically, as previously observed for other morphological and reproductive characters (Vicentini, 2007, 2016), our discriminant analysis based on few vegetative characters (leaf shape, leaf area and number of secondary veins) was able to assign samples to species (but not to subspecies) with high accuracy even for species not clearly recognized by visual inspection. These results highlight the importance of applying morphometric analysis based on digital images for cryptic species identification in the Amazon, especially because of the high plant diversity, the lack of specialists in many plant families and the many taxonomic uncertainties in herbarium collections. For example, the digitalization of all Amazonian herbarium specimens (ter Steege *et al.*, 2016) could enable the application of morphometric analysis for virtually any herbarium specimen image. Our results also demonstrate for the first time the power of NIR leaf-spectroscopy to discriminate species and subspecies with high accuracy in a species complex, reinforcing our species hypotheses and the use of this technique as a powerful alternative for specimen identification and species discovery in the Amazonian flora (Durgante *et al.*, 2013; Lang *et al.*, 2015; ter Steege *et al.*, 2016).

Regarding the role of niche evolution in the processes of lineage diversification and its importance for species delimitation, here we found that ecological niches are not conserved among sister or closely related species in the PGC. Niche conservatism is the process by which species tend to retain ancestral characteristics related to ecological specialization (Wiens & Graham, 2005). However, when in sympatry, closely related species may diverge ecologically to avoid or minimize resources overlap (Losos, 2008), as already reported by Vicentini (2007, 2016) for *Pagamea* spp. For instance, the sister species *P. spruceana* and *P.* m. *occulta* are found in sympatry (at least < 10 km apart) in the Upper Rio Negro region, but predominantly distributed in different parts of the flooding gradient (the former in long-term flooded areas and the second in less flooded habitats). It is likely that the micro-allopatry condition created by habitat preferences may have played a role in the diversification processes by reducing gene flow between these species, as suggested for *Protium subserratum* (Engl.) Engl. (Burseraceae) in the Peruvian Amazon (Fine *et al.*, 2005; Misiewicz & Fine, 2014). We suspect that some degree of gene flow may still exist because we found one specimen (*PRATA-1949*) morphologically similar to *P.* m. *occulta* in a population of *P. spruceana* (in the same habitat) and placed in the *P.* m. *occulta* clade in the RaxML and Exabayes (concatenated) analyses, but more closely related to *P. spruceana* in the Astral-II (coalescent) species tree. Considering that multispecies coalescent analysis incorporates the probability of ILS (Mirarab & Warnow, 2015;

Yang, 2015; Edwards *et al.*, 2016), it is more likely that hybridization, rather than ILS, explains the uncertain position of this sample. Although this question needs further investigation, here we accept *P. spruceana* and *P.* m. *occulta* as sister species even though they may not be completely reproductively isolated (but morphologically, ecologically and NIR spectrally different), as a potential case of sympatric speciation with strong selection by habitat.

In this study, we demonstrated the importance of disentangling species complexes for the detection of cryptic species diversity in the discovery and description of the Amazonian flora. Species delimitation has important consequences for ecological studies such as those concerning species distribution. For instance, the concept of 'hyperdominant species' (ter Steege *et al.*, 2013) should be carefully applied, especially in the case of species complexes, where taxonomic errors may inflate species distributions, consequently affecting conservation status and conservation efforts. We believe that the integration of genomic-based, morphological (including classificatory analysis from images), ecological and NIR spectroscopy analyses, combined with taxonomic expertise, may represent a great advance for taxonomic identification of herbarium specimens, for species delimitation and for the discovery of new species in the Amazon.

ACKNOWLEDGEMENTS

We thank the staff from Herbarium INPA, Laboratório de Evolução Aplicada (LEA-UFAM), Laboratório Temático de Biologia Molecular (LTBM-INPA) and Evolutionary Genetics Lab (MVZ-UC Berkeley), especially Lydia Smith, for help with DNA library preparation. We thank Fernando O. G. Figueiredo for the important tips and suggestions on Ecological Niche Modelling. We also thank the Federação das Organizações Indígenas do Rio Negro (FOIRN), Fundação Nacional do Índio (FUNAI) and Instituto Chico Mendes de Conservação da Biodiversidade (ICMBio) for permits to collect into Protected Areas, and all the people who helped during fieldwork. We also thank the Editor-in-Chief, the Associate Editor and two anonymous reviewers for their revisions and comments, which greatly improved the manuscript. This study was part of EMBP PhD thesis. Funding was provided by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), grant #478539/2011-8 to AV, fellowship to EMBP, Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM), grant #062.00205/2013 to EMBP and National Science Foundation (Division of Environmental Biology), award #1254214 to PVAF.

REFERENCES

Aberer AJ, Kobert K, Stamatakis A. 2014. ExaBayes: massively parallel Bayesian tree inference for the whole-genome era. *Molecular Biology and Evolution* **31:** 2553–2556.

Bremer B, Eriksson T. 2009. Time tree of Rubiaceae: phylogeny and dating the family, subfamilies, and tribes. *International Journal of Plant Sciences* **170:** 766–793.

Broennimann O, Fitzpatrick MC, Pearman PB, Petitpierre B, Pellissier L, Yoccoz NG, Thuiller W, Fortin MJ, Randin C, Zimmermann NE, Graham CH, Guisan A. 2012. Measuring ecological niche overlap from occurrence and spatial environmental data. *Global Ecology and Biogeography* **21:** 481–497.

Cardoso D, Särkinen T, Alexander S, Amorim AM, Bittrich V, Celis M, Daly DC, Fiaschi P, Funk VA, Giacomin LL, Goldenberg R, Heiden G, Iganci J, Kelloff CL, Knapp S, Lima HC, Machado AFP, Santos RM, Mello-Silva R, Michelangeli FA, Mitchell J, Moonlight P, Moraes PLR, Mori SA, Nunes SA, Pennington TD, Pirani JR, Prance GT, Queiroz LP, Rapini A, Riina R, Rincon CAV, Roque N, Shimizu G, Sobral M, Stehmann JR, Stevens WD, Taylor CM, Trovó M, van den Berg C, van der Werff H, Viana PL, Zartman CE, Forzza RC. 2017. Amazon plant diversity revealed by a taxonomically verified species list. *Proceedings of the National Academy of Sciences of the United States of America* **114:** 10695–10700.

Carstens BC, Pelletier TA, Reid NM, Satler JD. 2013. How to fail at species delimitation. *Molecular Ecology* **22:** 4369–4383.

Carstens BC, Satler JD. 2013. The carnivorous plant described as *Sarracenia alata* contains two cryptic species. *Biological Journal of the Linnean Society* **109:** 737–746.

Carstens B, Lemmon AR, Lemmon EM. 2012. The promises and pitfalls of next-generation sequencing data in phylogeography. *Systematic Biology* **61:** 713–715.

Caviedes-Solis IW, Bouzid NM, Banbury BL, Leaché AD. 2015. Uprooting phylogenetic uncertainty in coalescent species delimitation: a meta-analysis of empirical studies. *Current Zoology* **61:** 866–873.

Chen C, Durand E, Forbes F, François O. 2007. Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Notes* **7:** 747–756.

Chou J, Gupta A, Yaduvanshi S, Davidson R, Nute M, Mirarab S, Warnow T. 2015. A comparative study of SVDquartets and other coalescent-based species tree estimation methods. *BMC Genomics* **16:** 1–11.

Cortes C, Vapnik V. 1995. Support-vector networks. *Machine Learning* **20:** 273–297.

Costa-Silva GJ, Rodriguez MS, Roxo FF, Foresti F, Oliveira C. 2015. Using different methods to access the difficult task of delimiting species in a complex neotropical hyperdiverse group. *PLoS One* **10:** e0135075.

Cracraft J. 1983. Species concepts and speciation analysis. *Current Ornithology* **1:** 159–187.

de Queiroz K. 1999. The general lineage concept of species and the defining properties of the species category. In: Wilson RA, ed. *Species: new interdisciplinary essays*. Cambridge: MIT Press, 49–84.

de Queiroz K. 2007. Species concepts and species delimitation. *Systematic Biology* **56:** 879–886.

Domingos FM, Bosque RJ, Cassimiro J, Colli GR, Rodrigues MT, Santos MG, Beheregaray LB. 2014. Out of the deep: cryptic speciation in a Neotropical gecko (Squamata, Phyllodactylidae) revealed by species delimitation methods. *Molecular Phylogenetics and Evolution* **80:** 113–124.

Donoghue M. 1985. A critique of the biological species concept and recommendations for a phylogenetic alternative. *American Bryological and Lichenological Society* **88:** 172–181.

Durgante FM, Higuchi N, Almeida A, Vicentini A. 2013. Forest ecology and management species spectral signature: discriminating closely related plant species in the Amazon with near-infrared leaf-spectroscopy. *Forest Ecology and Management* **291:** 240–248.

Edwards SV, Xi Z, Janke A, Faircloth BC, McCormack JE, Glenn TC, Zhong B, Wu S, Lemmon EM, Lemmon AR, Leaché AD, Liu L, Davis CC. 2016. Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. *Molecular Phylogenetics and Evolution* **94:** 447–462.

Fine PV, Daly DC, Villa Muñoz G, Mesones I, Cameron KM. 2005. The contribution of edaphic heterogeneity to the evolution and diversity of Burseraceae trees in the western Amazon. *Evolution; International Journal of Organic Evolution* **59:** 1464–1478.

Fine PV, Zapata F, Daly DC. 2014. Investigating processes of neotropical rain forest tree diversification by examining the evolution and historical biogeography of the Protieae (Burseraceae). *Evolution; International Journal of Organic Evolution* **68:** 1988–2004.

Fujita MK, Leaché AD, Burbrink FT, McGuire JA, Moritz C. 2012. Coalescent-based species delimitation in an integrative taxonomy. *Trends in Ecology & Evolution* **27:** 480–488.

Garcia MG, Silva RS, Carniello MA, Veldman JW, Rossi AAB, Oliveira LO. 2011. Molecular evidence of cryptic speciation, historical range expansion, and recent intraspecific hybridization in the Neotropical seasonal forest tree *Cedrela fissilis* (Meliaceae). *Molecular Phylogenetics and Evolution* **61:** 639–649.

Grube M, Kroken S. 2000. Molecular approaches and the concept of species and species complexes in lichenized fungi. *Mycological Research* **104:** 1284–1294.

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W. 2004. Ten species in one: DNA barcoding reveals cryptic species in the Neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences of the United States of America* **101:** 14812–14817.

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

Hijmans RJ, Phillips S, Leathwick J, Elith J. 2017. *dismo: species distribution modeling. R package version 1.1–4*. https://CRAN.R-project.org/package=dismo

Iwata H, Ukai Y. 2002. SHAPE: a computer program package for quantitative evaluation of biological shapes based on elliptic Fourier descriptors. *The Journal of Heredity* **93:** 384–385.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30:** 772–780.

Kircher M, Sawyer S, Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research* **40:** e3.

Knowles LL, Carstens BC. 2007. Delimiting species without monophyletic gene trees. *Systematic Biology* **56:** 887–895.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* **22:** 568–576.

Kuhl PF, Giardina CR. 1982. Elliptic Fourier features of a closed contour. *Computer Graphics and Image Processing* **18:** 236–258.

Lang C, Costa FR, Camargo JL, Durgante FM, Vicentini A. 2015. Near infrared spectroscopy facilitates rapid identification of both young and mature Amazonian tree species. *PLoS One* **10:** e0134521.

Leaché AD, Fujita MK. 2010. Bayesian species delimitation in West African forest geckos (*Hemidactylus fasciatus*). *Proceedings of the Royal Society B* **277:** 3071–3077.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics (Oxford, England)* **25:** 2078–2079.

Losos JB. 2008. Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. *Ecology Letters* **11:** 995–1003.

Maddison WP. 1997. Gene trees in species trees. *Systematic Biology* **46:** 523–536.

Mayr E. 1992. A local flora and the biological species concept. *American Journal of Botany* **79:** 222–238.

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, Chang C-C, Lin C-C. 2017. *e1071: misc functions of the Department of Statistics, Probability Theory Group (formerly: E1071), TU Wien. R package version 1.6–8*. https://CRAN.R-project.org/package=e1071

Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols* **2010:** pdb.prot5448.

Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* **31:** 44–52.

Mishler BD, Brandon RN. 1987. Individuality, pluralism, and the phylogenetic species concept. *Biology and Philosophy* **2:** 397–414.

Misiewicz TM, Fine PV. 2014. Evidence for ecological divergence across a mosaic of soil types in an Amazonian tropical tree: *Protium subserratum* (Burseraceae). *Molecular Ecology* **23:** 2543–2558.

Muscarella R, Galante PJ, Soley-Guardia M, Boria RA, Kass JM, Uriarte M, Anderson RP. 2014. ENMeval: an R package for conducting spatially independent evaluations and estimating optimal model complexity for MAXENT ecological niche models. *Methods in Ecology and Evolution* **5:** 1198–1205.

Phillips SJ, Dudík M. 2008. Modeling of species distribution with Maxent: new extensions and a comprehensive evalutation. *Ecography* **31:** 161–175.

Piedra-Malagón EM, Albarrán-lara AL, Rull J, Piñero D, Sosa V. 2016. Using multiple sources of characters to delimit species in the genus *Crataegus* (Rosaceae): the case of the *Crataegus rosei* complex. *Systematics and Biodiversity* **14:** 244–260.

Prata EMB, Carvalho RB, Vicentini A. 2016. *Pagamea spruceana* (Rubiaceae, Gaertnereae), a new species from flooded white-sand forests in the Upper Rio Negro region, Brazil. *Phytotaxa* **269:** 186–192.

R Core Team. 2017. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Rannala B. 2015. The art and science of species delimitation. *Current Zoology* **61:** 846–853.

Rannala B, Yang Z. 2015. Efficient Bayesian species tree inference under the multi-species coalescent. *Systematic Biology* **66:** 823–842.

Sass C, Iles WJ, Barrett CF, Smith SY, Specht CD. 2016. Revisiting the Zingiberales: using multiplexed exon capture to resolve ancient and recent phylogenetic splits in a charismatic plant lineage. *PeerJ* **4:** e1584.

Singhal S. 2013. *De novo* transcriptomic analyses for non-model organisms: an evaluation of methods across a multi-species data set. *Molecular Ecology Resources* **13:** 403–416.

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

ter Steege H, Pitman NCA, Sabatier D, Baraloto C, Salomão RP, Guevara JE, Phillips OL, Castilho CV, Magnusson WE, Molino J-F, Monteagudo A, Nuñez Vargas P, Carlos Montero J, Feldpausch TR, Coronado ENH, Killeen TJ, Mostacedo B, Vasquez R, Assis RL, Terborgh J, Wittmann F, Andrade A, Laurance WF, Laurance SGW, Marimon BS, Marimon B-H, Guimarães Vieira IC, Amaral IL, Brienen R, Castellanos H, López DC, Duivenvoorden JF, Mogollón HF, de Almeida Matos FD, Dávila N, García-Villacorta R, Stevenson Diaz PR, Costa F, Emilio T, Levis C, Schietti J, Souza P, Alonso A, Dallmeier F, Duque Montoya AJ, Fernandez Piedade MT, Araujo-Murakami A, Arroyo L, Gribel R, Fine PVA, Peres CA, Toledo M, Gerardo AAC, Baker TR, Ceron C, Engel J, Henkel TW, Maas P, Petronelli P, Stropp J, Eugene Zartman C, Daly D, Neill D, Silveira M, Rios Paredes M, Chave J, de Andrade Lima D, Jorgensen PM, Fuentes A, Schoengart J, Cornejo Valverde F, Di Fiore A, Jimenez EM, Penuela Mora MC, Fernando Phillips J, Rivas G, van Andel TR, von Hildebrand P, Hoffman B, Zent EL, Malhi Y, Prieto A, Rudas A, Ruschell AR, Silva N, Vos V, Zent S, Oliveira AA, Cano Schutz A, Gonzales T, Nascimento MT, Ramirez-Angulo H, Sierra R, Tirado M, Umana Medina MN, van der Heijden G, Vela CIA, Vilanova Torre E, Vriesendorp C, Wang O, Young KR, Baider C, Balslev H, Ferreira C, Mesones I, Torres-Lezama A, Urrego Giraldo LE, Zagt R, Alexiades MN, Hernandez L, Huamantupa-Chuquimaco I, Milliken W, Palacios Cuenca W, Pauletto D, Valderrama Sandoval E, Valenzuela Gamarra L, Dexter KG, Feeley K, Lopez-Gonzalez G, Silman MR. 2013. Hyperdominance in the Amazonian tree flora. *Science* **342:** 325.

ter Steege H, Vaessen RW, Cárdenas-López D, Sabatier D, Antonelli A, de Oliveira SM, Pitman NC, Jørgensen PM, Salomão RP. 2016. The discovery of the Amazonian tree flora with an updated checklist of all known tree taxa. *Scientific Reports* **6:** 29549.

Sukumaran J, Knowles LL. 2017. Multispecies coalescent delimits structure, not species. *Proceedings of the National Academy of Sciences of the United States of America* **114:** 1607–1612.

Van Valen L. 1976. Ecological species, multispecies, and oaks. *Taxon* **25:** 233–239.

Vicentini A. 2007. Pagamea *Aubl. (Rubiaceae), from species to processes, building the bridge*. Unpublished D. Phil. Thesis, University of Missouri Saint Louis.

Vicentini A, van der Werff H, Nicolau S. 1999. Lauraceae. In: Ribeiro JELS, Hopkins MJG, Vicentini A, Sothers CA, Costa MAS, Brito JM, Solza MA, Martins LH, Lohmann LG, Assunção PACL, Pereira EC, Silva CF, Mesquita MR, Procópio LC, eds. *Flora da reserva Ducke: guia de identificação das plantas vasculares de uma floresta de terra-firme na Amazônia central*. Manaus: INPA, 150–179.

Vicentini A. 2016. The evolutionary history of *Pagamea* (Rubiaceae), a white-sand specialist lineage in Tropical South America. *Biotropica* **48:** 58–69.

Warren DL, Glor RE, Turelli M. 2008. Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution; International Journal of Organic Evolution* **62:** 2868–2883.

Wiens JJ, Graham CH. 2005. Niche conservatism: integrating evolution, ecology, and conservation biology. *Annual Review of Ecology, Evolution, and Systematics* **36:** 519–539.

Wilson E, Brown WL. 1953. The subspecies concept and its taxonomic application. *Systematic Biology* **2:** 97–111.

Yang Z. 2015. The BPP program for species tree estimation and species delimitation. *Current Zoology* **61:** 854–865.

Yang Z, Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences of the United States of America* **107:** 9264–9269.

Yang Z, Rannala B. 2014. Unguided species delimitation using DNA sequence data from multiple loci. *Molecular Biology and Evolution* **31:** 3125–3135.

Zhang C, Zhang DX, Zhu T, Yang Z. 2011. Evaluation of a Bayesian coalescent method of species delimitation. *Systematic Biology* **60:** 747–761.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Figure S1. Phylogenetic tree of the *Pagamea guianensis* species complex based on RaxML analysis of 431 loci. Node values are in likelihood bootstrap percentages. *P. coriacea* and *P. acrensis* are outgroups. The scale bar is in units of substitutions per site.
Figure S2. Species tree of the *Pagamea guianensis* species complex based on Astral-II analysis of 431 loci. Node values are in likelihood bootstrap percentages. *P. coriacea* and *P. acrensis* are outgroups. The scale bar is in coalescent units.
Figure S3. Phylogenetic tree of the *Pagamea guianensis* species complex based on Exabayes analysis of 431 loci. Node values are in Posterior Probability. *P. coriacea* and *P. acrensis* are outgroups. The scale bar is in units of substitutions per site.
Figure S4. Ecological Niche Models (ENM) generated with Maxent for the 14 species of the *Pagamea guianensis* species complex.
Figure S5. Taxonomic history of the *Pagamea guianensis* species complex.
Table S1. Information on the samples used for the phylogenetic and species tree analysis of the *Pagamea guianensis* complex. Taxa names are according to our proposed species and subspecies