Title
Disentangling the evolutionary history of a hyperdominant plant taxon in the Neotropics

Author
Damasco do Vale, Gabriel

Publication Date
2019

Peer reviewed|Thesis/dissertation

Disentangling the evolutionary history of a hyperdominant plant taxon in the Neotropics

By

Gabriel Damasco Do Vale

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Integrative Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Paul V.A. Fine, Chair

Professor Bruce G. Baldwin

Professor Todd E. Dawson

Professor Rosemary G. Gillespie

Spring 2019

Abstract

Disentangling the evolutionary history of a hyperdominant plant taxon in the Neotropics

By

Gabriel Damasco Do Vale

Doctor of Philosophy in Integrative Biology

University of California, Berkeley

Associate Professor Paul V.A. Fine, Chair

With the technological advent of bioinformatics, next-generation sequencing, and population genomics, recent studies can rely on high-throughput molecular data to investigate the evolution and diversification of Neotropical plant lineages. Here, we review one of the most widespread and dominant plant groups in the Neotropics based on a multidisciplinary approach. We integrate molecular data based on next-generation sequencing, morphological and functional data to unveil the evolutionary history of *Protium heptaphyllum* (Aubl.) Marchand, a plant taxon recently classified as one of the top hyperdominant trees in Amazonia. The results have direct implications for biodiversity conservation, taxonomy and systematics of tropical plants. The dissertation chapters are described below.

The first chapter aimed to perform the most comprehensive review ever done in a hyperdominant taxa, long considered to be a taxonomically difficult group. By using morphological, genomic, and functional data, we showed that *P. heptaphyllum sensu lato* represents eight separately evolving lineages warranting species status. In addition, most of these newly discovered lineages are rare and threatened; few if any of them could be considered hyperdominant on their own. There is an urgent need to improve the classification of hyperdominant taxa in order to avoid oversimplified assumptions regarding diversity and functional aspects of tropical regions.

The second chapter comprises a taxonomic review and a detailed description of *P. cordatum*. Molecular phylogeny indicates that populations of *P. cordatum* should not be classified as an intraspecific taxon within *Protium heptaphyllum* (Burseraceae), while morphology and near-infrared spectroscopy data provide additional support for the recognition of a separate entity. Species delimitation remains a challenge worldwide and this study demonstrates the importance of using multiple tools to characterize and distinguish plant species in highly diverse tropical regions.

In the third chapter, I aimed to generate a novel genomic resource for Burseraceae, a family globally recognized for producing resins and essential oils with medical properties and economic values. In this section, I provide the transcriptome assembly of *Protium copal*, a widespread aromatic tree in Central America, and describe the functional annotation of terpene biosynthetic genes. Most of the aromatic and non-aromatic properties of Burseraceae resins are composed by terpene and terpenoid chemicals. The identification of terpene genes will be relevant for understanding the synthesis of economically important chemicals in Burseraceae.

CHAPTER 1 Revisiting the hyperdominance of Neotropical tree species under a taxonomic, functional and evolutionary perspective.

Abstract Recent studies have reached high-impact conclusions about the diversity and functional ecology of Amazonian tree communities ever since large datasets from plot-inventory networks have been combined and analyzed together. These results suggest a phenomenon of hyperdominance in Amazonian tree communities based on the evidence that few species are common, and many are rare. Here, we aimed to review one hyperdominant taxa, long considered to be a taxonomically difficult group, called *Protium heptaphyllum* (Burseraceae). By using morphological, genomic, and functional data, we showed that *P. heptaphyllum sensu lato* represents eight separately evolving lineages warranting species status and most of these newly discovered lineages are geographically restricted; few if any of them could be considered hyperdominant on their own. In addition, functional trait data suggests that trees from each lineage would respond differently to environmental conditions, and some lineages are rare, with habitats experiencing deforestation. There is an urgent need to improve sampling and methods for species discovery in order to avoid oversimplified assumptions regarding diversity and functional aspects of tropical regions.

Keywords *Protium heptaphyllum*; taxonomy; species complex; cryptic morphology; Amazonia; Burseraceae; conservation biology

INTRODUCTION

Accurately defining species distributions is critical to understanding the processes responsible for both the generation and maintenance of biodiversity. The increasing availability and analysis of species distribution data through forest inventory networks (e.g. ATDN, RAINFOR, PPBio, DRYFLOR, 2ndFOR) has resulted in important predictions about species diversity (e.g. ter Steege *et al.* 2013, Banda-R *et al.* 2016, ter Stegee *et al.* 2018, Draper *et al.* 2019) and related ecosystem processes (e.g. Fauset *et al.* 2015, Rozendaal *et al.* 2019). Despite the scientific advances provided by the analysis of large dataset networks, the diversity of Amazon plants remains critically understudied, even for tree species (Feeley 2015).

Estimates regarding the total species richness of the Amazonian tree flora have been recently debated. The numbers vary from 6,727 (Cardoso *et al.* 2017, based on taxa from regional floras) to 10,071 species (ter Steege *et al.* 2018, based on plot-based checklists screened for taxonomic validity), but the actual richness is expected to reach over 16,000 species (ter Steege *et al.* 2016), given the large number of unsampled areas and the high diversity of extremely rare trees (ter Steege *et al.* 2013). Discrepancies between empirical and estimated values have at least one of three explanations: i) undersampled regions; ii) undescribed or undetermined taxa, collected but yet to be identified; and iii) the presence of species complexes or cryptic lineages for which subtle morphological distinctions may have been recognized but no genetic data have been included to help establish whether putative taxa are distinct. Although 1,068 new species have been added since ter Steege (2016) published the first check-list, there have been no systematic efforts to review the taxonomy of hyperdominant groups, especially those considered taxonomically difficult groups, morphospecies, or species complexes.

Potential species complexes are characterized by at least one of the following criteria (Bickford *et al.* 2007, Pinheiro *et al.* 2018): i) cryptic morphological variation, ii) recent evolutionary divergence and genetic introgression, iii) a large geographic range with often incomplete sampling, and iv) extensive historical taxonomic synonymy. Plant species complexes are fairly common in the Amazon region (e.g., Pennington and Lavin 2017, Prata *et al.* 2018) and based on the criteria above, we suggest that at least half of the 50 most hyperdominant trees (*sensu* ter Steege *et al.* 2013) potentially represent species complexes. If some or all of these hyperdominant species are actually species complexes that warrant treatment as multiple species or separately evolving taxa, this will have many important consequences. Resolution of such complexes would refine predictions about Amazonian species diversity, identify previously unknown and potentially threatened taxa and key habitats as new conservation priorities, and enhance the understanding of ecosystem processes and global change if newly described taxa exhibit differential responses to variation in abiotic and/or biotic conditions.

Here, we review a dominant and widespread plant species, *Protium heptaphyllum* (Aubl.) Marchand, by conducting extensive population-level sampling across its geographic range. *Protium heptaphyllum* was listed as the 12th most common tree species in Amazonia (ter Steege *et al.* 2013) and it also lives in the Cerrado and Atlantic forest of Brazil. *Protium* is one of the taxonomically most well studied genera in the Neotropics (Daly *et al.* 2010, Daly *et al.* 2012, Fine *et al.* 2014, Daly and Fine 2018) and it was recently classified as the second most dominant tree genus in the Amazon Basin (ter Steege *et al.* 2013). We combined morphological, genomic, and physiological data to answer the following questions: i) Does this hyperdominant tree represent a single species or should it be treated as multiple species? ii) What processes appear to have generated and maintained evolutionary divergence of any lineages within *P. heptaphyllum*? iii) How much variation in functional traits is found among such lineages? The answers to these questions have important implications in understanding the diversity, ecosystem processes and conservation of Amazonian forests.

MATERIALS AND METHODS

Field data collection. — We conducted multiple field sampling across the geographic range of *Protium heptaphyllum s.l.* in South America (Figure 1). In total, 39 populations of *P. heptaphyllum* were sampled, in addition to ten closely related outgroup species selected based on a molecular phylogeny of the Protieae tribe (Fine *et al.* 2014), i.e., *P. unifoliolatum*, *P. trifoliolatum*, *P. krukoffii*, *P. pillosum*, *P. widgrenii*, *P. icicariba*, *P. kleinii*, *P. brasiliense*, *P. ovatum,* and *P. dawsonii*. Outgroup taxa were sampled in the field when possible; otherwise, herbarium specimens were used. For each sample, we recorded latitude, longitude, and elevation with a GPS device. Leaf material for DNA extraction was dried right after collection using silica gel and subsequently stored in a -20ºC freezer. In addition, we performed leaf morphological measurements, extracted DNA, and collected functional trait data on the same plant specimens. Soil texture and fertility was analyzed from each population site. Morphological, ddRAD sequences, and functional trait data are provided in the supporting information. Voucher specimens were deposited at the New York Botanical Garden (NY) and the University Herbarium (UC).

Morphological measurements. — We generated a character matrix with 59 continuous and 98 discrete traits to investigate the variability of morphological characters in a multidimensional space. Non-informative characters and missing data were excluded. We used the R package

clustvarsel v.2.3.3 (Scrucca and Raftery 2014) to reduce the dimensionality of the data by selecting the set of principal components most useful for discrimination without a priori grouping information. Vegetative and reproductive traits were measured on 104 specimens of *P. heptaphyllum s.l.*. To test the hypothesis that *P. heptaphyllum s.l.* represents multiple lineages worthy of taxonomic recognition, we fit the number of morphological clusters using the normal mixture models (NMMs) implemented in the R package mclust v.5.0 (Scrucca *et al.* 2016). The Bayesian information criterion (BIC) was used to evaluate the best-fit number of morphological groups according to each NMM (Cadena *et al.* 2018).

DNA extraction and ddRAD library preparation. — We extracted high-quality genomic DNA from 415 samples of *P. heptaphyllum s.l.* widely distributed throughout the Amazon, Atlantic Forest, and Cerrado, and six outgroup species. DNA was extracted from ca. 100 mg of leaf tissue preserved in silica or from herbarium specimens when silica-dried leaves were not available. Extractions followed an updated version of the DNEasy Plant mini kit protocol (Qiagen, Crawley, U.K.). Double-digest RAD-seq libraries were prepared for high-throughput sequencing following Peterson *et al.* (2012) and DNA was digested with *Sph*I-HF and *Eco*RI-HF enzymes. DNA libraries were sequenced on five lanes of an Illumina HiSeq 4000 at the University of Berkeley QB3 facility.

ddRAD data analysis. — We assembled the ddRAD-seq reads using the software ipyrad version 0.6.17 (Eaton and Overcast 2016), in order to generate two *de novo* assemblies with different levels of sample sizes and missing data. First, we used a clustering threshold of sequence similarity set to 0.85 and retained a data set including all loci shared by at least four samples. All other ipyrad parameters were set to default. In total, 23,922 loci and 39,770 SNPs were recovered from 415 samples (dataset 1). Although the average sample coverage was high (385 loci per sample), 56% of samples had high levels of missing data. We generated another assembly using a higher cluster threshold of sequence similarity (0.90) and including all loci shared by at least 10 samples. The low-missing-data assembly resulted in 6,234 total filtered loci and 25,027 SNPs from 285 samples, and 21% missing data (dataset 2).

Genetic differentiation. — To estimate the number of genetic clusters within *P. heptaphyllum s.l.* without a priori assumptions of individual assignments, we used the Bayesian clustering algorithm STRUCTURE (Pritchard *et al.* 2000, Evanno *et al.* 2005). We ran 10 replicates at each value of K for 500,000 generations with a burn-in of 50,000. $K = 3$ was preferred with the ΔK method and had the highest log P(X|K). Those genetic clusters corresponded with the populations located in three major biogeographic regions: i) the Amazon biome, ii) ecotone zones between the Amazon and neighboring biomes, and iii) a large geographic region encompassing the Cerrado and Atlantic Forest biomes (Figure 1). In order to identify finer patterns of genetic structure, we ran a separate analysis within the three identified groups using similar parameter settings. We used Structure Harvester version 0.6.94 to compare alternative values of K based on the log probability (log P(X|K)) and the ΔK statistic (Evanno *et al.* 2005). Pairwise F_{ST} was calculated among populations using Weir and Cockerham (1984) equations implemented in the WCfst function from the hierfstat R package (Goudet 2005). D_{xy} was calculated as described in Nei (1987), with a custom R script slightly modified from the genet.dist function in hierfstat. D_{xy} measures absolute genetic distance, while F_{ST} measures genetic differentiation among populations relative to within populations (Nei 1987). F_{ST} values under 0.19 (or 0.20) are usually interpreted as lacking significant genetic structuring or population subdivision. At the same time, F_{ST} values above 0.3 are seen as moderate

population structuring. F_{ST} values above 0.5 are normally considered to be associated with strong population subdivision.

Divergence time and population size estimates. — The demographic parameters were implemented in the Generalized Phylogenetic Coalescent Sampler (G-PhoCS version 1.2.3, Gronau *et al.* 2011), a Bayesian method for inferring divergence times and effective population sizes from genome sequences. First, we generated a *de novo* assembly with 1,971 loci and 43 samples selected from populations within the 10 different clusters defined by STRUCTURE. In the MCMC runs, we used the gamma distribution with $\alpha = 1.0$ and $\beta = 10{,}000$ for the mutation-scaled population sizes and divergence times, and a gamma distribution with $\alpha = 0.002$ and $\beta = 0.00001$ for the mutation-scaled migration rates. Each Markov Chain was run for 100,000 burn-in iterations, after which parameter values were sampled for 500,000 iterations every 50 iterations. Convergence was inspected manually for each run. Parameter calibration in the probabilistic model of G-PhoCS is scaled by mutation rate μ. We assumed an average mutation rate of $\mu = 7.2 \times 10^{-5}$ mutations per site per generation given by $\tau = T\mu/g$, where τ is the parameter tau, g is the average generation time (in years), and T is the absolute divergence time (in years). We used a divergence time of 5 Ma (95% CI: 3.8–11.1 Ma) for the clade including *P. heptaphyllum* and close relative outgroups based on Fine *et al.* (2014) and the generation time of five years was based on an estimation of flowering time data observed in the field. Effective population sizes are given by $\theta = 4N_e\mu$, where θ is parameter theta and N_e is the absolute effective population size, and the migration rates are given by $M = m/\mu$, where m is the probability of migration across two given populations.

Phylogenetic reconstruction. — To reconstruct phylogenetic relationships among individuals we took multiple approaches. Trees were inferred from concatenated data sets using RAxML version 8.2.10 (Stamatakis 2014) and ExaBayes version 1.5 (Aberer *et al.* 2014) and the dataset consisted of 285 samples and 6,234 loci. For RAxML, maximum likelihood phylogenies were inferred with 1000 bootstraps using the GTRGAMMA substitution model. For ExaBayes analysis, we ran two Metropolis-coupling replicates with four coupled-chains (each with three heated chains) for 1 × 106 MCMC generations, sampled every 500 generations. Estimated sample size (ESS) for all parameters and branch lengths were summarized with the postProcParam tool and confirmed as sufficiently sampled after analyzing in Tracer v.1.6 (Aberer *et al.* 2014). Finally, we generated a majority-rule consensus tree with the Exabayes 'consense' tool after a 25% burn-in. Each tree was rooted by the clade composed of *P. unifoliolatum*, *P. trifoliolatum*, *P. krukoffii*, *P. brasiliense*, *P. icicariba*, and *P. widgrenii*.

Species tree inference and species delimitation. — We inferred species trees using the coalescent-based methods SVDquartet (Chifman and Kubatko 2014) conducted in PAUP 4a157 (http://phylosolutions.com/paup-test/), and BPP version 3.4 (Flouri *et al.* 2018). The data was scored with different populations set as partitions using the A01 analysis (Yang 2015). For SVDquartet, we used default settings except for the species tree option and conducted 1000 bootstrap replicates. The posterior distribution was independent of different starting species trees and the topology was consistent when comparing results of multiple runs. We used the A10 analysis in BPP version 3.4 to test species delimitation within the *P. heptaphyllum s.l.* group using a fixed species tree previously inferred by SVDquartet and BPP. We assigned equal prior probabilities to all species delimitation models (1 to 10 species) and all runs were sampled every 50 generations for 10,000 samples with a burn-in of 2,000. We evaluated convergence by

comparing results of replicate runs. In BPP, we assigned equal probabilities for the rooted trees and set the inverse-gamma priors $\theta \sim IG(2, 50)$ for all θs and $\tau \sim IG(4, 50)$ for the age of the root ($\tau 0$) according to the demographic inference previously performed by G-PhoCS analysis.

Functional traits. — In order to understand the physiological responses of distinct *P. heptaphyllum s.l.* populations across different climatic and environmental conditions, we measured multiple functional characters including leaf, wood, and chemical traits. The traits measured were specific leaf area (SLA), leaf nitrogen content, leaf carbon content, leaf stable carbon isotope (13C) composition, leaf secondary chemistry, leaf chlorophyll content, leaf stomatal density, wood vessel size, and wood vessel density. These traits have been reported as important drivers of the leaf economic spectrum of plants (Wright *et al.* 2004). Details about each functional trait is available in Table S1.

RESULTS

Multi-evidence species delimitation. Morphological analysis suggests that at least six distinct groups can be resolved based on normal mixture models (NMM) assuming 1-10 distinct morphological groups. The NMM results indicated a best-fit model with six distinct morphological groups (BIC-value -4,500, Figure S1). Nonetheless, models set for seven and eight species had similar but slightly lower BIC-values (-4,532 and -4,528, respectively). The STRUCTURE analysis of the genomic data recognized three major clusters that represent a broad pattern of genetic differentiation among major biogeographic regions in South America: i) Central-Western Amazonia, ii) Northeastern Amazonia, and iii) Cerrado and Atlantic Forest domains). However, a second STRUCTURE run within the major clusters revealed a finer pattern of population and lineage subdivision and identified a total of nine groups strongly supported according to the ΔK method (Figure 1) and log P(X|K). RAxML and ExaBayes species trees are compatible with the STRUCTURE results and all nine genetic clusters are monophyletic. SVDquartet and BPP (A01 analysis) also strongly support the monophyly of nine genetic clusters (posterior probability >0.99). Finally, the BPP A10 species delimitation analysis supported a ten-species model with posterior probabilities higher than 0.75 and a nine-species model with posterior probabilities higher than 0.20 regardless of the prior settings used. Based on the total set of evidence, we conclude that *P. heptaphyllum s.l.* represents multiple evolutionary lineages that warrant treatment as separate species. Many of these lineages are geographically restricted. When overlaying the geographic ranges of these lineages onto the density of *Protium heptaphyllum s.l.* from the ATDN dataset (Figure 1), it is clear that the individual trees in different areas are likely to be assigned to many different new taxa. Although it is possible that one (or more) of these lineages could still be considered "hyperdominant" (i.e. one of the most common 227 trees in Amazonia that account for half of all stems), this will have to be tested with morphological or molecular analysis of voucher specimens from the ATDN plots.

Demographic history and genetic differentiation. — G-PhoCS estimates of divergence time and effective population sizes suggest that *P. heptaphyllum s.l.* lineages experienced distinct evolutionary histories of demographic expansion. *P. heptaphyllum* populations diverged from the outgroup clade ca. 5 Ma after an abrupt increase in population size (Figure 2). According to divergence time results, the earliest-diverging *P. heptaphyllum* lineages evolved in the Amazon basin and currently occur in white-sand forests and temporarily flooded habitats. On the other hand, late-diverging lineages diversified throughout the tropical rain forests and seasonally dry

habitats from Central Brazil and the Atlantic coast very recently, ca. 0.6 Ma. Besides representing multiple distinct species in Amazonia, *P. heptaphyllum s.l.* shows contrasting values of current effective population sizes within its lineages. In terms of genetic differentiation, F_{ST} and D_{xy} metrics indicate high similarity and potential gene flow among relatively recently diverged lineages from the Cerrado and Atlantic Forest (Figure 3). As lineages become less phylogenetically related and spatially distant, F_{ST} and D_{xy} values increase, suggesting high genetic differentiation between Amazonian populations and more recently-diverged lineages from Central Brazil and the Atlantic coast.

Functional variation and environmental structure. — Functional traits within *P. heptaphyllum s.l.* showed very different patterns in different populations (Figure 4, Figure S2). In addition, populations located at local habitat boundaries also displayed distinct functional strategies. Some lineages had very narrow trait variation while others showed a broad range. While populations in Amazonia usually had consistent and less variable traits, lineages from Central Brazil Cerrado and Atlantic Forest often showed greater intraspecific trait variation. For instance, *P. heptaphyllum* populations from the Atlantic Forest were commonly sampled across forest-savanna ecotones (e.g. rainforest and *restinga* transects, populations with code MUS and PON, respectively) and leaf and habit traits usually match the local soil fertility and water availability (i.e. stunted trees or shrubs less than 2 m tall inhabiting dry and poor soils known as coastal *restingas,* and large 30 m trees occupying nutrient-rich and moisture soils in the neighboring forest habitat). The soil cation exchange capacity (CEC), measured as a proxy for fertility and water availability, had a strong relationship with functional trait strategies (Figure 4). Independently of their phylogeography, populations inhabiting nutrient-poor and dry soils had higher SLA (small and thick leaves), higher 13C ratios (lower photosynthetic activity and higher water use efficiency) and slightly higher stomatal density. Thus, we found that *P. heptaphyllum s.l.* exhibited a complex and variable profile of trait characters and the clear functional differences among several groups suggest important differential responses to heterogeneous environments.

DISCUSSION

Based on an extensive geographic sampling and on morphological, ddRADseq and functional data, we have identified several distinct evolutionary lineages within *P. heptaphyllum*. We discard the hypothesis that *P. heptaphyllum s.l.* is a single species and instead find evidence for recognizing at least eight independently evolving species (or lineages) (*sensu* de Queiroz 2007). These eight lineages are strongly supported and represent genetically differentiated populations with distinct morphological and functional traits. In terms of demographic history, older lineages tend to be habitat specialists and less morphologically and functionally variable. In contrast, recently evolved lineages have recently colonized new areas and frequently experience gene flow across very large geographic distances, and they are morphologically and functionally more variable. Our results have important implications for taxonomy, biodiversity conservation, and ecosystem functioning. Even though a large number of new species have been recently described in the Neotropics (Baker *et al.* 2017, Cardoso *et al.* 2017, ter Steege *et al.* 2018), very few studies have attempted to resolve morphologically challenging taxa within species complexes (e.g. Maas *et al.* 2015, Prata *et al.* 2018). In the section bellow, we discuss the implications of our results in the context of: i) revisiting the concept of hyperdominance for Amazonian trees and improving richness and diversity estimates, ii) understanding diversification within dominant tropical lineages, iii) refining ecosystem functioning predictions and iv) conservation of rare and threatened taxa.

Implications for the hyperdominance phenomenon. — The fact that communities often harbor a small group of demographically abundant species in addition to a much larger number of rare species is not a recent discovery (e.g. Preston 1948, as cited in Draper *et al.* 2019). This pattern, also called a species oligarchy and hyperdominance, was first reported for the Amazon forest in the early 2000's (Pitman *et al.* 2001, Macía and Svenning 2005) and again at the pan-Amazonian scale by ter Steege *et al.* (2013), based on data from the Amazon Tree Diversity Network (ATDN, http://atdn.myspecies.info/), the largest tree community datasets ever compiled in the tropics. Hyperdominant species have captured the imagination of many tropical ecologists. First, they have been thought to be more likely to be correctly identified than rare species (Baker *et al.* 2017) allowing for people to use them as proxies for ecosystem-wide function. For example, Fauset *et al.* (2015) claimed that hyperdominant species were responsible for half of carbon storage and productivity in the Amazon. Second, ecologists have noted that hyperdominants have important shared demographic properties – they often have large geographic ranges but are only dominant in one or two regions of the Amazon basin and are often habitat specialists (ter Steege *et al.* 2013). In contrast, our results suggest that the hyperdominant taxon *P. heptaphyllum s.l.* actually consists of several lineages warranting recognition as new species that have very distinct geographic ranges and include several that are rare or threatened. We wonder if similar conclusions could be reached with many (or most) of the other hyperdominant species, which are also thought to be members of species complexes (e.g., *Irartea deltoidea* Henderson 1995; *Eschweilera coriacea*, Mori *et al.* 2017).

Taxonomic relevance. — *Protium* is one of the best studied plant groups in the Neotropics (Daly *et al.* 2012, Fine *et al.* 2014). Currently, the genus consists of approximately 200 species and their taxonomic treatment has been studied by a collaborative team of taxonomists and evolutionary biologists. *Protium* has a wide geographic range of specimen sampling and genomic data are available for several species (Fine *et al.* 2014, Afonso *et al.* 2018, Damasco *et al.* in review). In addition, species descriptions in *Protium* are consistently founded on both morphological and molecular phylogenetic evidence. This gives us high confidence in our results presented here, which cannot yet be said for groups that have not been subjected to intensive systematic study (but often include diverse and abundant trees in the Amazon basin (i.e., Myrtaceae, Lauraceae).

In *P. heptaphyllum s.l.*, our results confirmed that morphological traits overlap at some level, but genetic and chemical data suggest that this hyperdominant and widespread group represents distinct evolutionary and independently adapted lineages that diverged over a million years ago. Here, we showed that a multidisciplinary effort and relatively short time investment on a species complex (3-5 years) has yielded the discovery of several new lineages warranting species status. Taxonomic updates within *P. heptaphyllum s.l.* are in progress to be published (e.g. *Protium cordatum* Huber *sensu* Damasco *et al.* in press) and detailed descriptions of new species are in preparation as part of a taxonomic revision (Damasco in prep.).

Several diverse botanical families are also well studied in the Neotropics (e.g., Annonaceae, Sapotaceae, Lecytidaceae, Rubiaceae, Chrysobalanaceae, Fabaceae), but frequently cited for containing a few lineages with incomplete genetic divergence that need to be studied in great detail (e.g., *Inga*, Pennington and Lavin 2015). We acknowledge that the field of taxonomy is dynamic (ter Steege *et al.* 2018) in the sense that classification and species names are likely to change as new studies are performed. Besides that, the methods applied to describe, reestablish, and invalidate taxonomic entities can be quite inconsistent among taxonomists and experts, as it varies

accordingly to several species' concepts and definitions (Zachos 2016, Garnett and Christidis 2017). The same disagreement of inconsistent methods in taxonomy has been the focus of recent debates about the tree species richness in Amazonia. The discussion raised by Cardoso *et al.* (2017) and ter Steege *et al.* (2018) is very valuable to tropical science, but both studies agree that there is still much work to be done in order to improve taxonomy and increase the pace of species discovery (Baker *et al.* 2017).

Refining the understanding of trait variation and functional response. — Aside from comprising separately morphological and genetic lineages (six distinct lineages for the Central/West Amazon Basin and Northeast Amazonian ecotone regions), *P. heptaphyllum s.l.* also showed substantial variation in leaf, wood, and chemical traits. Our secondary metabolite results showing little variation within populations yet strongly divergent chemical profiles in sister lineages are consistent with what has been reported in a recent study on chemical diversity in the genus *Protium*, (Salazar *et al.* 2018) as well as a few other tropical plant lineages (Richards *et al.* 2007, Endara *et al.* 2015, Sedio *et al.* 2017).

Since the advent of functional trait networks (e.g. TRY Plant Trait Database, Kattge *et al.* 2011), understanding how plant species behave in terms of their physiological performance over large scales has led to important predictions related to future global change and land-use scenarios (Laurance *et al.* 2016, Ewers *et al.* 2017, Esquivel-Muelbert *et al.* 2019). Yet, the accuracy of global vegetation models commonly used to address these questions depends on the quality of species determination and proper standardization of specimens across different plots (Baker *et al.* 2017). Considering the importance of the Amazon region for the global climate and carbon cycle (Saleska *et al.* 2003, Mahli *et al.* 2008), it is extremely important to devote substantial effort to investigating the taxonomy of hyperdominant plant taxa. According to Fauset *et al.* (2015), carbon stocks in the most diverse regions on Earth are concentrated in remarkably few dominant species. However, if some or most of the hyperdominant taxa actually represent multiple evolutionary entities, as in *P. heptaphyllum s.l.*, a larger fraction of Amazonian tree species would contribute proportionally more to carbon storage and cycling than described by Fauset *et al.* (2015). Thus, a taxonomic review of dominant tree lineages would greatly benefit the understanding of tree community physiological responses to global change.

Understanding diversification in dominant lineages. — The scenario of some hyperdominant or oligarchic taxa representing multiple diverged lineages is intriguing and relevant for understanding the processes of diversification in the Amazonian flora. One might hypothesize that large population sizes may be associated with higher diversification rates due to the process of population expansion followed by specialization into different habitats. Therefore, dominant lineages would have higher chances to speciate via habitat specialization and generate large clades than non-dominant lineages. Habitat specialization is considered to have evolved in many tropical plant groups (Fine *et al.* 2004, 2006) and many different tree genera have become specialized to contrasting environments (Fine and Baraloto 2016).

According to our results, *P. heptaphyllum s.l.* diverged from its common ancestors around five million years ago and diversified first in the Amazon region followed by an abrupt increase in population size. Many lineages subsequently became specialized in white-sand habitats, seasonally-flooded forests in the Rio Negro Basin (*Igapó* forests), and floodplain forests (*Baixio* forests). However, more recently, lineages dispersed into neighboring floristic domains (e.g.

Cerrado and Atlantic Forest) and ecotone areas, colonizing regions where the annual precipitation is currently lower and seasonal. These populations found in Central and Coastal Brazil are genetically very similar to each other despite the large geographical distances among them. Besides that, these populations are functionally more variable than the early-diverging lineages in the Amazon. In contrast, older lineages from the Amazon basin were found to be less morphologically plastic and more genetically isolated in terms of gene flow. White-sand ecosystems in Amazonia are characterized by a patchy and geographically disjunct distribution (Adeney *et al.* 2016) that could have inhibited dispersal of habitat-specialist populations.

Conservation Implications. — We show that at least four newly discovered lineages, including two resurrected species, are geographically restricted, demographically rare and endemic to white-sand vegetation in the Amazon (e.g. *P. cordatum sensu* Damasco *et al.* in press*, P.* "tucuruiense"*, P. angustifolium* Swart. (in a new sense)*,* and *P.* "reticuliflorum"). Further studies aiming to review potential species complexes among hyperdominant species potentially could identify other threatened lineages warranting taxonomic recognition and indicate new geographic areas for conservation priority. While many studies have relied on datasets compiled by plot-inventory networks, more effort should be focused on increasing the pace of taxonomic research in the Neotropics (Baker *et al.* 2017). Deforestation rates in Amazonia are likely to increase in the next few years and yet one third of the total number of tree species are likely still undescribed or undiscovered (ter Steege *et al.* 2013).

We believe that reports that roughly half of all biomass and carbon storage in one of the most diverse areas on Earth belong to a very small group of "hyperdominant" species may be misinforming policy makers and stakeholders. Based on our integrative review of *P. heptaphyllum s.l.*, we showed that genetic diversity and functional responses to environmental gradients are much greater than expected by the hyperdominance principle. Our results highlight that within a single hyperdominant taxon exists several lineages warranting recognition as distinct species that include great functional and genetic diversity, including a few that are very rare and extremely threatened by predicted land use scenarios. Simplistic assumptions on biogeochemical processes and ecosystems services should not be taken at face value based on extrapolations from a few hyperdominant tree species.

REFERENCES

Aberer, A.J., Kobert, K. and Stamatakis, A., 2014. ExaBayes: massively parallel Bayesian tree inference for the whole-genome era. Molecular biology and evolution, 31(10), pp.2553-2556.

Adeney, J.M., Christensen, N.L., Vicentini, A. and Cohn-Haft, M., 2016. White-sand ecosystems in Amazonia. Biotropica, 48(1), pp.7-23.

Afonso, L.F., Amaral, D., Uliano-Silva, M., Torres, A.L.Q., Simas, D.R. and de Freitas Rebelo, M., 2018. First Draft Genome of a Brazilian Atlantic Rainforest Burseraceae reveals commercially-promising genes involved in terpenic oleoresins synthesis. BioRxiv, p.467720.

Baker, T.R., Pennington, R.T., Dexter, K.G., Fine, P.V., Fortune-Hopkins, H., Honorio, E.N., Huamantupa-Chuquimaco, I., Klitgård, B.B., Lewis, G.P., de Lima, H.C. and Ashton, P.,

2017. Maximising synergy among tropical plant systematists, ecologists, and evolutionary biologists. Trends in ecology & evolution, 32(4), pp.258-267.

Banda-R, K., Delgado-Salinas, A. & Dexter, K.G. *et al.*, 2016. Plant diversity patterns in neotropical dry forests and their conservation implications. Science, 353, 1383-1387.

Bickford, D., Lohman, D.J., Sodhi, N.S., Ng, P.K., Meier, R., Winker, K., Ingram, K.K. and Das, I., 2007. Cryptic species as a window on diversity and conservation. Trends in ecology & evolution, 22(3), pp.148-155.

Cadena, C.D., Zapata, F., and Jiménez, I. 2018. Issues and perspectives in species delimitation using phenotypic data: Atlantean evolution in Darwin's finches. Syst. Biol. 67: 181–194.

Cardoso, D., Särkinen, T., Alexander, S., Amorim, A.M., Bittrich, V., Celis, M., Daly, D.C., Fiaschi, P., Funk, V.A., Giacomin, L.L. and Goldenberg, R., 2017. Amazon plant diversity revealed by a taxonomically verified species list. Proceedings of the National Academy of Sciences, 114(40), pp.10695-10700.

Chifman, J. and Kubatko, L., 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics, 30(23), pp.3317-3324.

Daly, D.C. and Fine, P.V., 2018. Generic limits re-visited and an updated sectional classification for Protium (tribe Protieae). Studies in Neotropical Burseraceae XXV. Brittonia, 70(4), pp.418-426.

Daly, D.C., Harley, M.M., Martínez-Habibe, M.C. and Weeks, A., 2010. Burseraceae. In Flowering Plants. Eudicots (pp. 76-104). Springer, Berlin, Heidelberg.

Daly, D.C.D.B., Fine, P.V.A. and Martínez-Habibe, M.C., 2012. Burseraceae: a model for studying the Amazon flora. Rodriguésia, 63(1), pp.021-030.

de Queiroz, K., 2007. Species concepts and species delimitation. Systematic biology, 56(6), pp.879-886.

Draper, F.C., Asner, G.P., Honorio Coronado, E.N., Baker, T.R., García-Villacorta, R., Pitman, N.C., Fine, P.V., Phillips, O.L., Zárate Gómez, R., Amasifuén Guerra, C.A. and Flores Arévalo, M., 2019. Dominant tree species drive beta diversity patterns in Western Amazonia. Ecology, p.e02636.

Eaton, D.A.R. and Overcast, I., 2016. ipyrad: interactive assembly and analysis of RADseq data sets.

Endara, M.J., Weinhold, A., Cox, J.E., Wiggins, N.L., Coley, P.D. and Kursar, T.A., 2015. Divergent evolution in antiherbivore defences within species complexes at a single Amazonian site. Journal of Ecology, 103(5), pp.1107-1118.

Esquivel-Muelbert, A., Baker, T.R., Dexter, K.G., Lewis, *S.L.*, Brienen, R.J., Feldpausch, T.R., Lloyd, J., Monteagudo-Mendoza, A., Arroyo, L., Álvarez-Dávila, E. and Higuchi, N., 2019. Compositional response of Amazon forests to climate change. Global change biology, 25(1), pp.39-56.

Evanno, G., Regnaut, S. and Goudet, J., 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular ecology, 14(8), pp.2611-2620.

Ewers, R.M., Andrade, A., Laurance, S.G., Camargo, J.L., Lovejoy, T.E. and Laurance, W.F., 2017. Predicted trajectories of tree community change in Amazonian rainforest fragments. Ecography, 40(1), pp.26-35.

Fauset, S., Johnson, M.O., Gloor, M., Baker, T.R., Monteagudo, A., Brienen, R.J., Feldpausch, T.R., Lopez-Gonzalez, G., Malhi, Y., Ter Steege, H. and Pitman, N.C., 2015. Hyperdominance in Amazonian forest carbon cycling. Nature communications, 6, p.6857.

Feeley, K.J. and Silman, M.R., 2011. The data void in modeling current and future distributions of tropical species. Global Change Biology, 17(1), pp.626-630.

Fine, P.V. and Baraloto, C., 2016. Habitat endemism in white-sand forests: insights into the mechanisms of lineage diversification and community assembly of the Neotropical flora. Biotropica, 48(1), pp.24-33.

Fine, P.V., Mesones, I. and Coley, P.D., 2004. Herbivores promote habitat specialization by trees in Amazonian forests. science, 305(5684), pp.663-665.

Fine, P.V., Miller, Z.J., Mesones, I., Irazuzta, S., Appel, H.M., Stevens, M.H.H., Sääksjärvi, I., Schultz, J.C. and Coley, P.D., 2006. The growth–defense trade-off and habitat specialization by plants in Amazonian forests. Ecology, 87(sp7), pp.S150-S162.

Fine, P.V., Zapata, F. and Daly, D.C., 2014. Investigating processes of neotropical rain forest tree diversification by examining the evolution and historical biogeography of the Protieae (Burseraceae). Evolution, 68(7), pp.1988-2004.

Flouri, T., Jiao, X., Rannala, B. and Yang, Z., 2018. Species tree inference with bpp using genomic sequences and the multispecies coalescent. Molecular biology and evolution, 35(10), pp.2585-2593.

Garnett, S.T. and Christidis, L., 2017. Taxonomy anarchy hampers conservation. Nature News, 546(7656), p.25.

Goudet, J., Jombart, T. and Goudet, M.J., 2015. Package 'hierfstat'. R package version 0.04-22. Retrieved from http://www. r-project. org, http://github. com/jgx65/hierfstat.

Gronau, I., Hubisz, M.J., Gulko, B., Danko, C.G. and Siepel, A., 2011. Bayesian inference of ancient human demography from individual genome sequences. Nature genetics, 43(10), p.1031.

Henderson, A., Bernal, R. and Galeano-Garces, G., 1997. Field guide to the palms of the Americas. Princeton University Press.

Kattge, J., Diaz, S., Lavorel, S., Prentice, I.C., Leadley, P., Bönisch, G., Garnier, E., Westoby, M., Reich, P.B., Wright, I.J. and Cornelissen, J.H.C., 2011. TRY–a global database of plant traits. Global change biology, 17(9), pp.2905-2935.

Laurance, W.F., Camargo, J.L., Fearnside, P.M., Lovejoy, T.E., Williamson, G.B., Mesquita, R.C., Meyer, C.F., Bobrowiec, P.E. and Laurance, S.G., 2016. An Amazonian forest and its fragments as a laboratory of global change. In Interactions between biosphere, atmosphere and human land use in the Amazon Basin (pp. 407-440). Springer, Berlin, Heidelberg.

Maas, P.J.M., Westra, L.Y.T., Guerrero, S.A., Lobão, A.Q., Scharf, U., Zamora, N.A. and Erkens, R.H.J., 2015. Confronting a morphological nightmare: revision of the Neotropical genus Guatteria (Annonaceae). Blumea-Biodiversity, Evolution and Biogeography of Plants, 60(1-2), pp.1-219.

Macía, M.J. and Svenning, J.C., 2005. Oligarchic dominance in western Amazonian plant communities. Journal of Tropical Ecology, 21(6), pp.613-626.

Malhi, Y., Roberts, J.T., Betts, R.A., Killeen, T.J., Li, W. and Nobre, C.A., 2008. Climate change, deforestation, and the fate of the Amazon. science, 319(5860), pp.169-172.

Mori, S.A., Kiernan, E.A., Smith, N.P., Kelly, L.M., Huang, Y.Y., Prance, G.T. and Thiers, B.M., 2017. Observations on the phytogeography of the Lecythidaceae clade (Brazil nut family). Guy L. Nesom.

Nei, M. and Li, W.H., 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences, 76(10), pp.5269-5273.

Pennington, R.T. and Lavin, M., 2016. The contrasting nature of woody plant species in different neotropical forest biomes reflects differences in ecological stability. New Phytologist, 210(1), pp.25-37.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S. and Hoekstra, H.E., 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PloS one, 7(5), p.e37135.

Pinheiro, F., Dantas-Queiroz, M.V. and Palma-Silva, C., 2018. Plant Species Complexes as Models to Understand Speciation and Evolution: A Review of South American Studies. Critical reviews in plant sciences, 37(1), pp.54-80.

Pitman, N.C., Terborgh, J.W., Silman, M.R., Núñez, P., Neill, D.A., Cerón, C.E., Palacios, W.A. and Aulestia, M., 2001. Dominance and distribution of tree species in upper Amazonian terra firme forests. Ecology, 82(8), pp.2101-2117.

Prata, E.M., Sass, C., Rodrigues, D.P., Domingos, F.M., Specht, C.D., Damasco, G., Ribas, C.C., Fine, P.V. and Vicentini, A., 2018. Towards integrative taxonomy in Neotropical botany: disentangling the Pagamea guianensis species complex (Rubiaceae). Botanical Journal of the Linnean Society, 188(2), pp.213-231.

Preston, F.W., 1948. The commonness, and rarity, of species. Ecology, 29(3), pp.254-283.

Pritchard, J.K., Stephens, M. and Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics, 155(2), pp.945-959.

Richards, L.A., Dyer, L.A., Forister, M.L., Smilanich, A.M., Dodson, C.D., Leonard, M.D. and Jeffrey, C.S., 2015. Phytochemical diversity drives plant–insect community diversity. Proceedings of the National Academy of Sciences, 112(35), pp.10973-10978.

Rozendaal, D.M., Bongers, F., Aide, T.M., Alvarez-Dávila, E., Ascarrunz, N., Balvanera, P., Becknell, J.M., Bentos, T.V., Brancalion, P.H., Cabral, G.A. and Calvo-Rodriguez, S., 2019. Biodiversity recovery of Neotropical secondary forests. Science advances, 5(3), p.eaau3114.

Salazar, D., Lokvam, J., Mesones, I., Vásquez Pilco, M., Ayarza Zuniga, J.M. and de Valpine, P., 2018. Origin and maintenance of chemical diversity in a species-rich tropical tree lineage. Nat. Ecol. Evol, 2, pp.983-990.

Saleska, S.R., Miller, S.D., Matross, D.M., Goulden, M.L., Wofsy, S.C., Da Rocha, H.R., De Camargo, P.B., Crill, P., Daube, B.C., De Freitas, H.C. and Hutyra, L., 2003. Carbon in Amazon forests: unexpected seasonal fluxes and disturbance-induced losses. Science, 302(5650), pp.1554-1557.

Scrucca, L. and Raftery, A.E., 2014. clustvarsel: A package implementing variable selection for model-based clustering in R. arXiv preprint arXiv:1411.0606.

Scrucca, L., Fop, M., Murphy, T.B. and Raftery, A.E., 2016. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R journal, 8(1), p.289.

Sedio, B.E., Rojas Echeverri, J.C., Boya, P., Cristopher, A. and Wright, S.J., 2017. Sources of variation in foliar secondary chemistry in a tropical forest tree community. Ecology, 98(3), pp.616-623.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, 30(9), pp.1312-1313.

ter Steege, H., de Oliveira, S.M., Pitman, N.C., Sabatier, D., Antonelli, A., Andino, J.E.G., Aymard, G.A. and Salomão, R.P., 2019. Towards a dynamic list of Amazonian tree species. Scientific reports, 9(1), p.3501.

ter Steege, H., Pitman, N.C., Sabatier, D., Baraloto, C., Salomão, R.P., Guevara, J.E., Phillips, O.L., Castilho, C.V., Magnusson, W.E., Molino, J.F. and Monteagudo, A., 2013. Hyperdominance in the Amazonian tree flora. Science, 342(6156), p.1243092.

ter Steege, H., Vaessen, R.W., Cárdenas-López, D., Sabatier, D., Antonelli, A., De Oliveira, S.M., Pitman, N.C., Jørgensen, P.M. and Salomão, R.P., 2016. The discovery of the Amazonian tree flora with an updated checklist of all known tree taxa. Scientific Reports, 6, p.29549.

Weir, B.S. and Cockerham, C.C., 1984. Estimating F-statistics for the analysis of population structure. evolution, 38(6), pp.1358-1370.

Wright, I.J., Reich, P.B., Westoby, M., Ackerly, D.D., Baruch, Z., Bongers, F., Cavender-Bares, J., Chapin, T., Cornelissen, J.H., Diemer, M. and Flexas, J., 2004. The worldwide leaf economics spectrum. Nature, 428(6985), p.821.

Yang, Z., 2015. The BPP program for species tree estimation and species delimitation. Current Zoology, 61(5), pp.854-865.

Zachos, F.E., 2016. Species concepts in biology (Vol. 801). Cham, Switzerland: Springer.

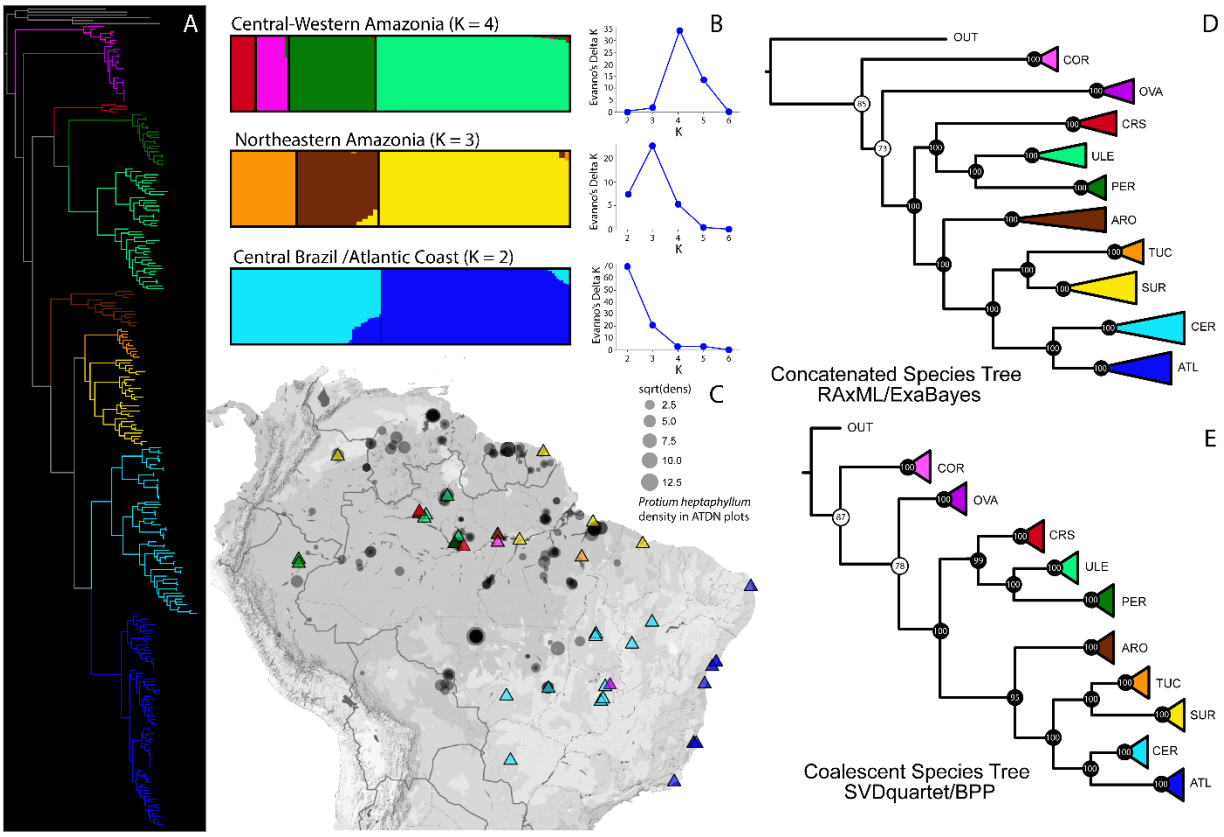

Figure 1. Phylogenetic reconstruction and population genetic structure of *Protium heptaphyllum s.l.* (Aubl.) Marchand. (A) Maximum likelihood (RAxML) phylogeny of *P. heptaphyllum* inferred with 1000 bootstrap replicates using the GTRGAMMA substitution model. The dataset consists of 285 samples and 6,234 loci. (B) STRUCTURE analysis identified three main clusters that encompasses the Central/West and North/East of Amazonia, and the Central and Atlantic Coast of Brazil. The bar graphs represent a subsequent STRUCTURE run within each geographic group based on a combination of 10 replicate runs. Colors indicate the assignment of each individual to a particular genetically similar cluster. The number of clusters within each geographic region were identified based on the ΔK statistics. (C) The location of each sampled population is displayed on the map (triangle symbols). Colors indicate the population assignment to a particular genetic cluster. Circle symbols corresponds to the plot location where *P. heptaphyllum s.l.* was sampled by the Amazon Tree Diversity Network (ATDN). Circle sizes are equivalent to the square root of density values per hectare in each of the ATDN plots. (D-E) Species tree inference based on concatenated data matrices ran in RAxML and ExaBayes

software and based on the multispecies coalescent model ran in SVDquartet and BPP software, respectively. Node values represent bootstrap support and posterior probabilities for each clade. White circles correspond to support values that are not identical between RAxML bootstrapping and ExaBayes posterior probabilities. In this case, the lower support value was displayed in the figure.

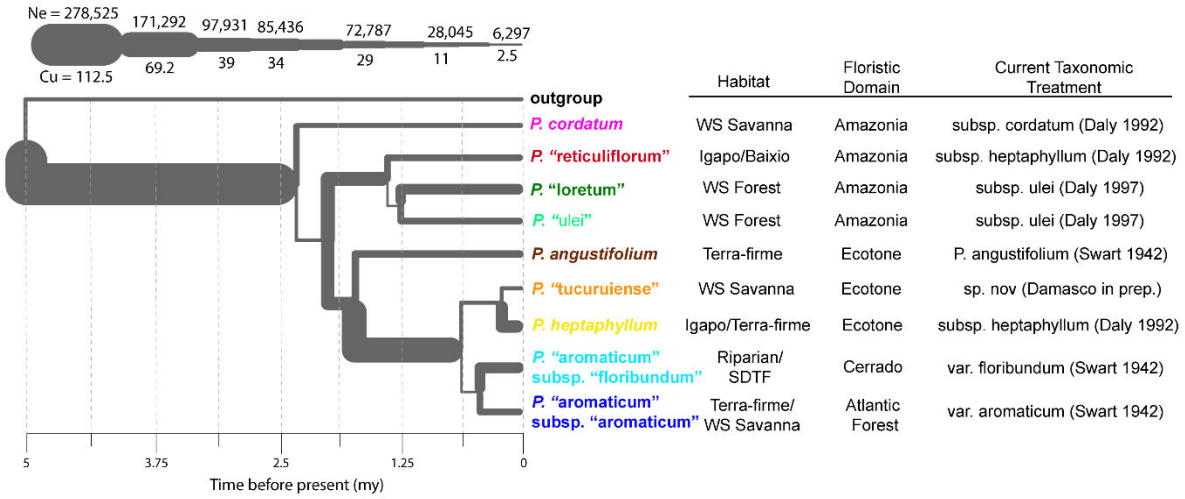

Figure 2. Historical demography and divergence estimate of *Protium heptaphyllum s.l.* populations. Effective population sizes (Ne) based on coalescent units (Cu), and divergence times estimated using the Generalized Phylogenetic Coalescent Sampler (G-PhoCS). The branch ranges (or width) correspond to 95% Bayesian credible intervals aggregated across three runs. Tip labels and colors represent the genetic population assignment based on STRUCTURE analysis. The plant names in quotes are informal, and they will be published in a separate taxonomic review of *P. heptaphyllum s.l.*. On the right, the table displays the habitat type, floristic domain, and the current taxonomic treatment for each lineage. The floristic domain "Ecotone" corresponds to populations located in transitional areas between the Amazonian floristic domains and neighboring biome areas (e.g. Dry Forests and Grasslands in Colombia).

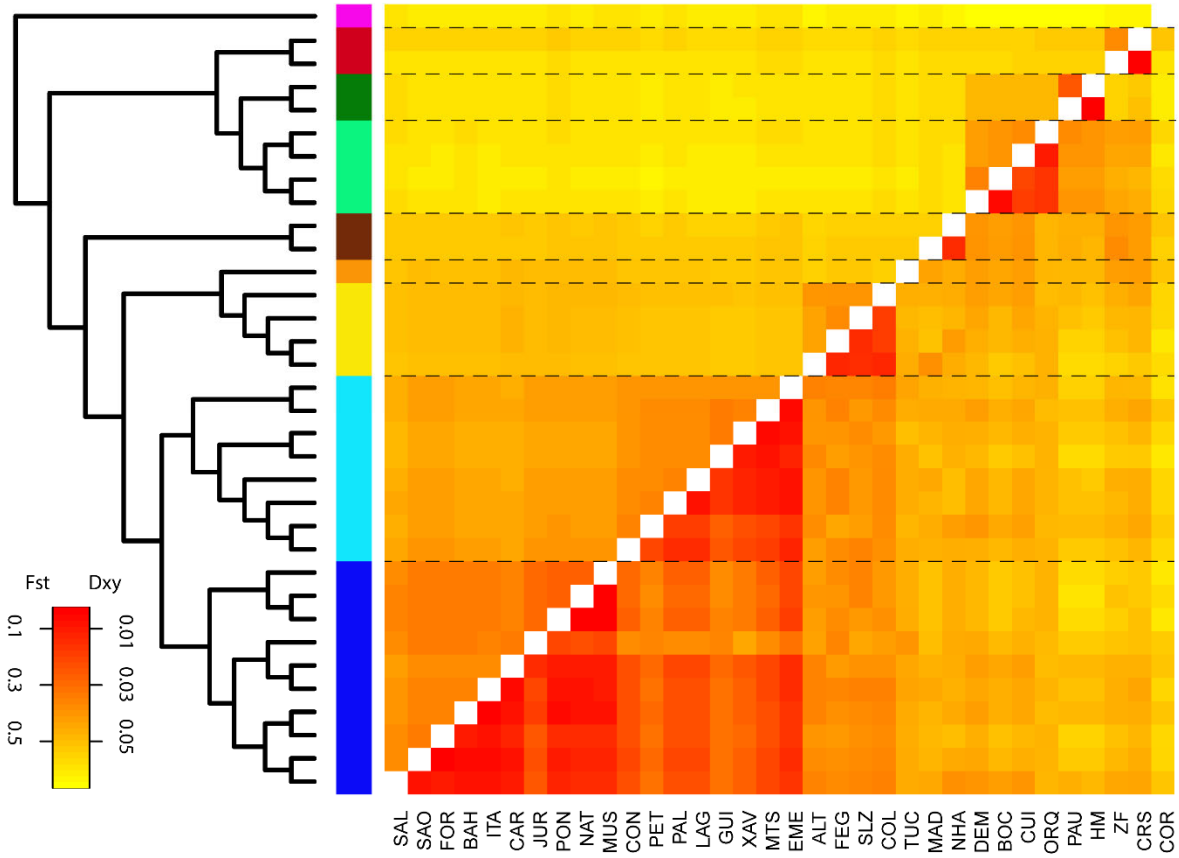

Figure 3. Pair-wise genetic distances among *P. heptaphyllum s.l.* populations based on two metrics: F_{ST} (lower matrix), a relative measure of genetic differentiation within and between pair-wise populations, and D_{xy} (upper matrix), an absolute measure of genetic divergence between pair-wise populations. Colored bars correspond to populations assignment based on the STRUCTURE analysis.

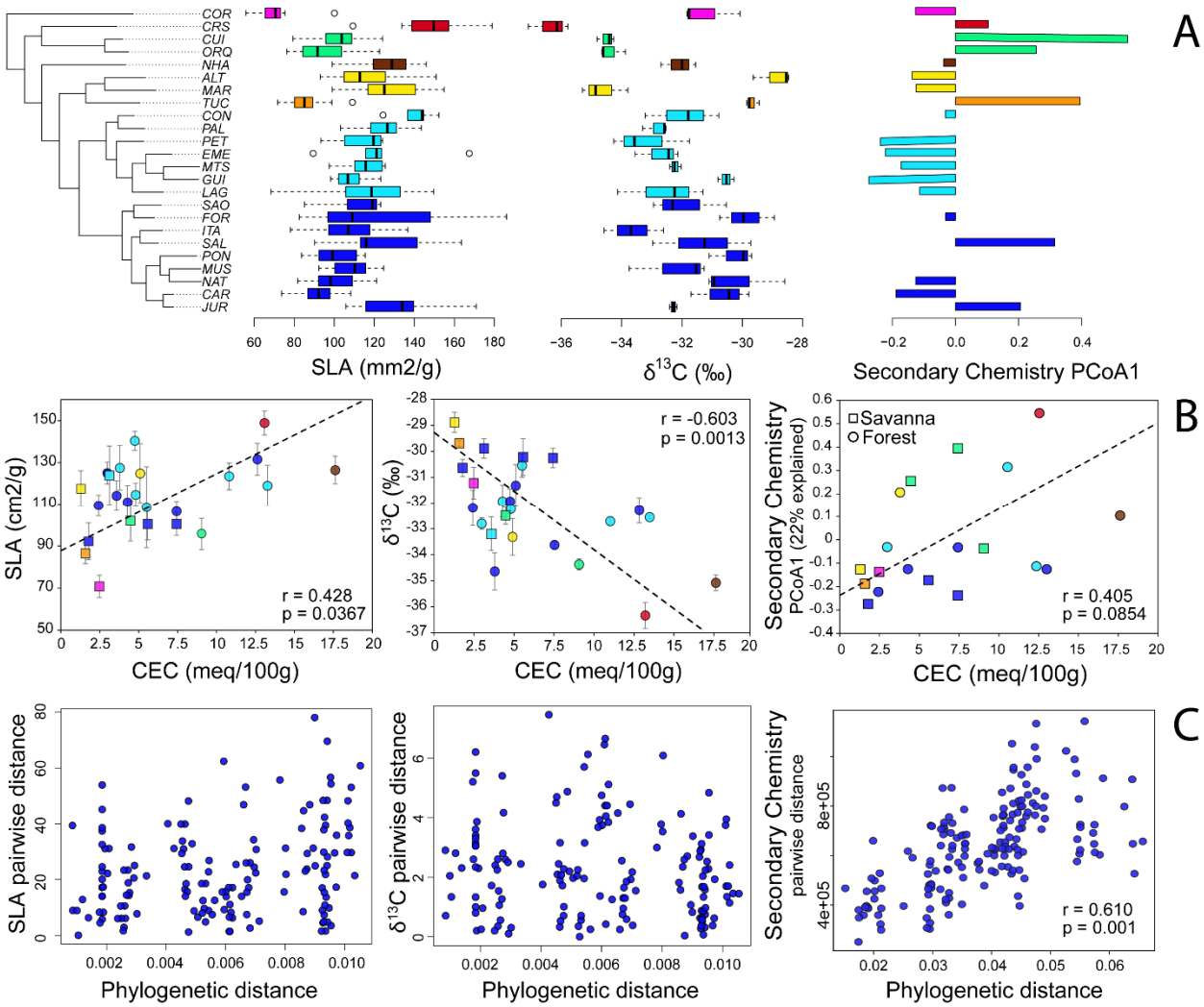

Figure 4. High trait lability among different populations of *Protium heptaphyllum s.l.* (A) Trait value variation for specific leaf area (SLA), stable isotope composition ($\delta^{13}C$), and secondary chemistry represented by the Principal Coordinate Analysis (PCoA) first axis. PCoA 1 axis explained 22% of the whole secondary chemical variation among *P. heptaphyllum s.l.* populations. (B) Correlation between the functional traits SLA, $\delta^{13}C$, and secondary chemistry and the soil cation exchange capacity (CEC). CEC is an approximate measurement of the soil nutrient and moisture content and high CEC values are associated with higher nutrient and water availability. (C) Correlation between trait pair-wise distances and the phylogenetic distances among populations of *P. heptaphyllum s.l.*.

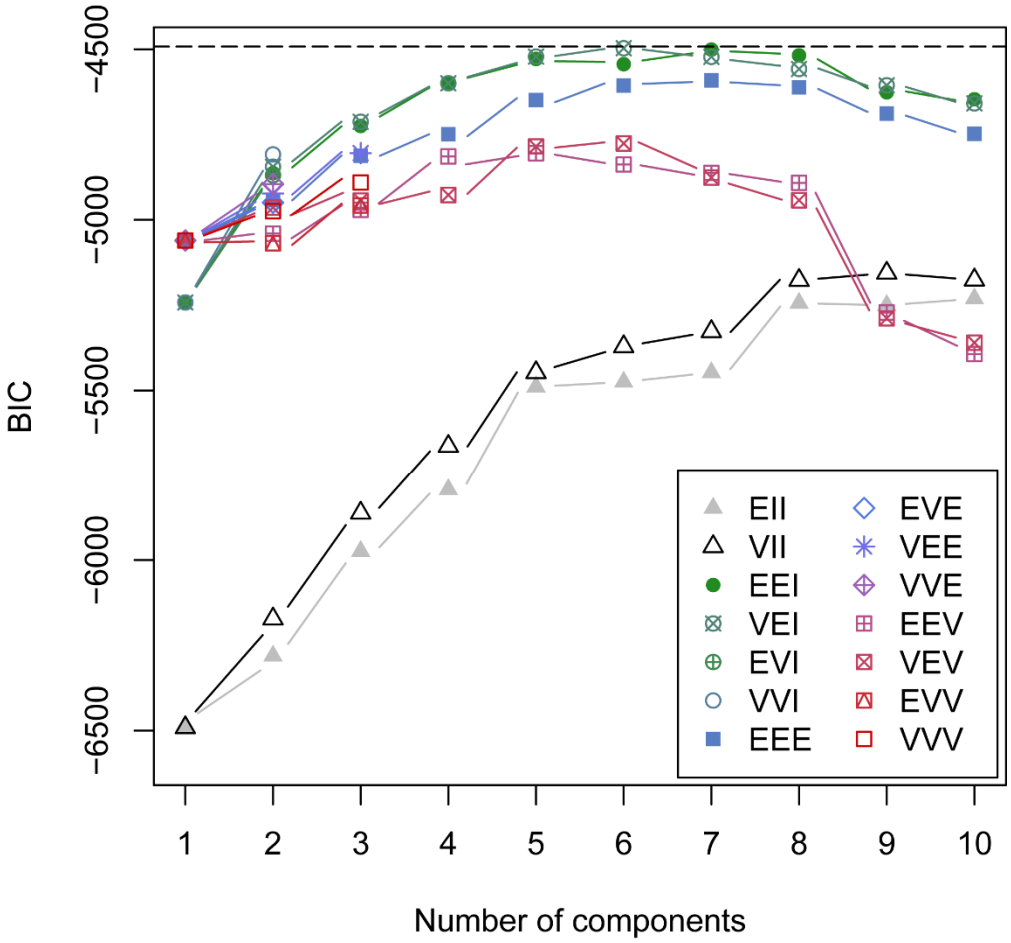

Figure S1. Morphological data strongly supported hypotheses that there are multiple distinct groups within *Protium heptaphyllum s.l.* The graph shows the support for normal mixture models (NMM) assuming 1-10 distinct morphological groups and 12 model parameterizations. The two models with highest support based on BIC values assumed six and seven distinct morphological groups (models EEI and VEI).

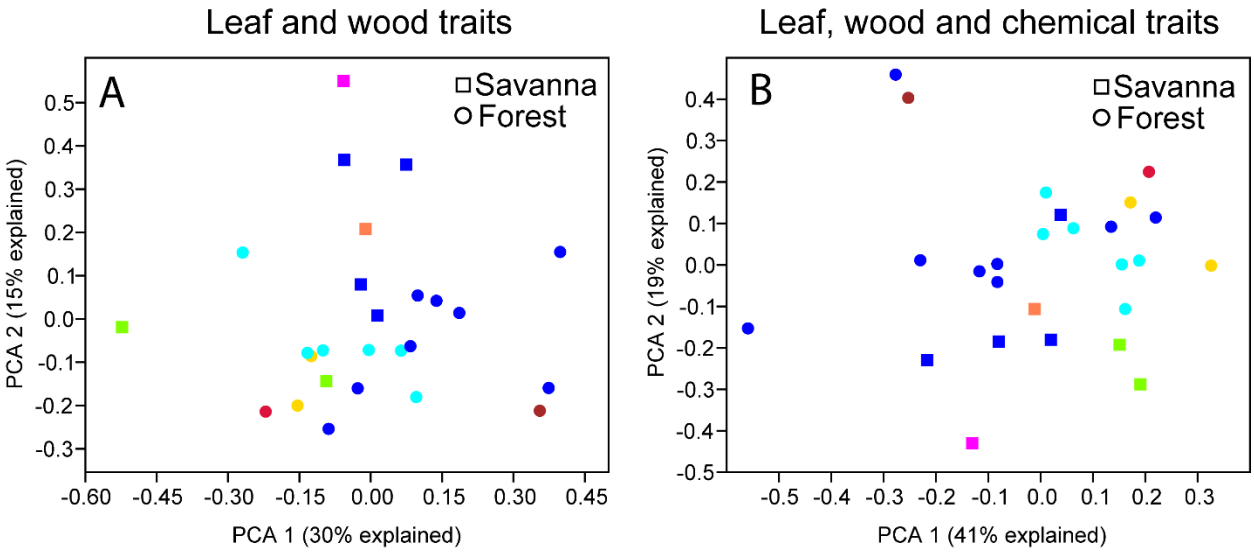

Figure S2. Multivariate analysis of leaf, wood, and chemical functional traits from different populations of *Protium heptaphyllum s.l.* (A) Principal Component Analysis (PCA) of leaf (specific leaf area, stable isotope composition, leaf nitrogen content, chlorophyll content, stomatal density) and wood traits (vessel diameter, vessel length, vessel density). (B) PCA for all the leaf and wood traits mentioned above, but with the addition of 4,618 secondary chemical variables. The addition of secondary chemistry into the PCA analysis increased the variation explained by the first and second PCA axis.

Table S1. Functional trait description for different populations of *Protium heptaphyllum s.l.*. Leaf and wood functional traits are displayed bellow. Information regarding the number of replicates, trait variation range, and proxy for trait strategies is described in detail.

Functional trait	Unit	Group	Replicates per population	Trait range (min - max)	Strategy
Specific Leaf Area (SLA)	cm^2g^{-1}	Leaf	12	70.78 – 148.88	Investment in photosynthetic capacity versus resource conservation and herbivory defense.
Stable Isotope Composition ($\delta^{13}C$)	‰	Leaf	3	-36.33 – -28.88	Intrinsic water use efficiency (WUE) and water-related limitations.
Leaf Nitrogen Content (N)	%	Leaf	3	1.21 – 1.88	Photosynthetic capacity.
Chlorophyll Content Index (CCI)	%	Leaf	12	19.19 – 29.58	Photosynthetic capacity.
Stomatal Density (SD)	mm^{-2}	Leaf	5	140.2 – 263.5	Assuming similar values of stomatal size, indicates higher CO_2 uptake.
Vessel diameter	µm	Wood	5	26.87 – 38.63	Hydraulic efficiency.
Vessel length	µm	Wood	5	71.17 – 131.42	Hydraulic efficiency and vulnerability to cavitation.
Vessel density	mm^{-2}	Wood	5	179 – 264.7	Hydraulic efficiency.

CHAPTER 2 Reestablishment of *Protium cordatum* (Burseraceae) based on integrative taxonomy

Abstract Species delimitation remains a challenge worldwide, but especially in biodiversity hotspots such as the Amazon. Here, we use an integrative taxonomic approach that combines data from morphology, phylogenomics, and leaf spectroscopy to clarify the species limits within the *Protium heptaphyllum* species complex, which includes subsp. *cordatum*, subsp. *heptaphyllum*, and subsp. *ulei*. Molecular phylogeny indicates that populations of subsp. *cordatum* do not belong to the P. heptaphyllum clade, while morphology and near-infrared spectroscopy data provide additional support for the recognition of a separate taxon. *Protium cordatum* (Burseraceae) is reinstated at species rank and described in detail. As circumscribed here, *P. cordatum* is endemic to white-sand savannas located in the Faro and Tucuruí Districts, Pará State, Brazil, whereas *P. heptaphyllum* is a dominant and widespread plant lineage found in Amazonia, the Cerrado, and the Brazilian Atlantic Forest. We present an identification key to *P. cordatum* and closely related lineages and a detailed taxonomic description of *P. cordatum*, including habitat and distribution, a list and images of diagnostic features. This study demonstrates the importance of using multiple tools to characterize and distinguish plant species in highly diverse tropical regions.

Keywords Amazon; campina; ddRAD; plant systematics; *Protium heptaphyllum*; near-infrared spectroscopy; white-sand forest

INTRODUCTION

The efficiency and accuracy of taxonomy have greatly increased with new technologies and analytical advances (Bik, 2017). Many species have recently been described based on different kinds of data that range from comparative morphology and anatomy to phylogeography, population genetics, and functional ecology (Dayrat, 2005; Schlick-Steiner *et al.*, 2010). The potential for such integrative taxonomic approaches has not yet been fully embraced in botany, particularly in the tropics, where biodiversity studies are especially needed. In highly diverse forests such as the Amazon, morphological overlap among closely related lineages complicates taxonomic delimitation. However, new data from molecular phylogenetics and DNA barcoding are contributing to the delimitation of species (Kress *et al.*, 2005; Gonzalez *et al.*, 2009). Furthermore, high-throughput genome sequencing has enhanced the resolution of population phylogeography and species delimitation analysis (e.g., Leaché *et al.*, 2014; Fišer *et al.*, 2018), leading to more accurate species limits.

In addition to molecular phylogenetics, Fourier-transformed near-infrared spectroscopy (FT-NIR) has demonstrated great potential for discriminating tropical plant species (Durgante *et al.*, 2013; Féret *and* Asner, 2013; Lang *et al.*, 2017). The spectrometer generates an absorbance response that is a function of the chemical composition and internal anatomy of leaves, which has been shown to be conserved within populations of the same species in a large study with more than 1000 tropical tree species (Asner *et al.*, 2014). Analyzing the spectral signatures of near-infrared reflectance of herbarium specimens can allow the discrimination of morphotypes even when there is a high phenotypic overlap among samples, and informative vegetative and reproductive traits are unknown (Durgante *et al.*, 2013; Prata *et al.*, 2018).

Protium (Burseraceae) has undergone rapid diversification in the Neotropics (Fine *et al.*, 2014) and includes many cryptic species. A good example is *Protium heptaphyllum* (Aubl.) Marchand, one of the most dominant trees in the Neotropics (Ter Steege *et al.*, 2013). *Protium heptaphyllum sensu lato* (*s.l.*) is also one of the most widespread plant taxa in the Neotropics and inhabits different biomes and related ecosystems (i.e., Amazonia – terra-firme forest and white-sand ecosystems; Cerrado and Pantanal – gallery forests and seasonally dry tropical forests; Guiana Shield – rocky savannas; and Brazilian Atlantic Forests – coastal white-sands and rain forests). The morphology of *P. heptaphyllum* can be quite variable, which may result from acclimation or adaptive divergence of populations inhabiting different soil types or climatic niches (i.e., Fine *et al.*, 2013). For instance, individuals found in sandy and nutrient-poor soils appear as shrubs or stunted trees with bifurcated trunks and coriaceous leaves, while individuals that inhabit more nutrient-rich clay soils can be tall canopy trees with chartaceous leaves.

Many infraspecific taxa have been assigned to *P. heptaphyllum* since Jean Baptiste C. Fusée Aublet first published *Icica heptaphylla* as part of the "Histoire des plantes de la Guiane françoise" in 1775 (Swart, 1942). Currently, three valid subspecies have been recognized in the *P. heptaphyllum* species complex (subsp. *cordatum*, subsp. *heptaphyllum*, subsp. *ulei*) yet no consensus has been reached regarding the degree to which phenotypic variation in the species complex corresponds to taxonomic entities (Daly, 1992).

Protium heptaphyllum subsp. *cordatum* was first described as *Protium cordatum* (at species rank) by Huber (1909) and subsequently treated as an infraspecific taxon within *P. heptaphyllum* due to quantitative overlap in measurements of the reproductive characters (Daly, 1992). Here, we use a multidisciplinary approach that combines next-generation sequencing, morphological analyses, and NIR spectral data and sample 24 populations across the geographic range of the *P. heptaphyllum* species complex to evaluate species limits within the group. Since taxonomy has moved towards being an integrative science, we believe that formal species should represent evolutionarily diverged populations that: (1) form highly supported monophyletic clades according to molecular evidence, (2) have limited gene flow among closely related lineages or sister groups, and (3) exhibit conserved morphological features that enable their recognition. We present evidence here that subsp. *cordatum* must be reinstated at species rank, as treated by Huber (1909).

MATERIALS AND METHODS

Taxon sampling. — We sampled 8 individuals of *P. cordatum* at the lectotype locality (Faro, Pará, Brazil) as well as 23 individuals of *P. heptaphyllum s.l.* from throughout its range (Fig. 1A). Since *P. cordatum* was previously treated as a subspecies of *P. heptaphyllum* (Daly, 1992), we also sampled 8 individuals that co-occurred with *P. cordatum* in the type locality (P. *cordatum* shrubs inhabiting the white-sand savanna and *P. heptaphyllum* adult trees inhabiting the adjacent forest; Fig. 1B). We aimed to test if these morphologically distinct populations from adjacent habitats represented genetically diverged populations. The morphological, molecular and spectral data were collected from the same samples. In addition, ten closely related outgroup species were selected based on a molecular phylogeny of the Protieae tribe (Fine *et al.*, 2014), i.e., *P. brasiliense*, *P. dawsonii*, *P. icicariba*, *P. kleinii*, *P. krukoffii*, *P. ovatum*, *P. pillosum*, *P. trifoliolatum*, *P. unifoliolatum*, and *P. widgrenii*.

Morphological analyses. — First, we generated a character matrix with 59 continuous and 98 discrete traits to examine the morphological variability in multidimensional space. Non-informative characters and missing data were excluded. We used the R package clustvarsel v.2.3.3 (Scrucca *et.* Raftery, 2014) to reduce the dimensionality of the data by selecting the set of principal components most useful for discrimination without a priori information about groups. Vegetative traits were measured on 56 specimens of *P. heptaphyllum s.l.* and 12 specimens of *P. cordatum*. Reproductive traits were measured on 23 specimens of *P. heptaphyllum* and 12 specimens of *P. cordatum*. Some specimens of the latter did not bear flowers or fruits during the sampling period. Therefore, descriptions of reproductive structures (e.g., corolla length, flower density and petal indumenta) were obtained from herbarium specimens collected in the same biogeographic domain (e.g., Cerrado, Amazonia, etc.).

To test the hypothesis that *P. cordatum* does not belong within *P. heptaphyllum* as an infraspecific taxon, we fit the number of morphological clusters using the normal mixture models (NMMs) implemented in the R package mclust v.5.0 (Scrucca *et al.*, 2016). The Bayesian information criterion (BIC; Schwarz, 1978) was used to evaluate the best-fit number of morphological groups according to each NMM (Cadena *et al.*, 2018). The vegetative characters with high loading values in the principal component analysis (PCA) were (1) leaf petiole length, (2) the maximum number of leaflets, (3) plant height and (4) specific leaf area. The reproductive traits with high loading PCA values were (1) flower density per inflorescence, (2) corolla length, and (3) petal indumenta density.

DNA library preparation. — We extracted high-quality genomic DNA from 8 samples of *P. cordatum*, 23 samples of *P. heptaphyllum s.l.* widely distributed throughout the Amazon, Atlantic Forest, and the Cerrado, and 10 outgroup species. DNA was extracted from ca. 100 mg of leaf tissue preserved in silica or from herbarium specimens when silica-dried leaves were not available. Extractions followed a modified version of the DNEasy Plant mini kit protocol (Qiagen, Crawley, U.K.). Double-digest RAD-seq libraries were prepared for high-throughput sequencing following Peterson *et al.* (2012). Detailed information on the library preparation procedures is available as Appendix S1. DNA was digested with SphI-HF and EcoRI-HF enzymes. DNA libraries were sequenced on five lanes of an Illumina HiSeq 4000 at the University of Berkeley QB3 facility.

Assembly and phylogenetic analysis. — We used a bioinformatics pipeline implemented in custom Perl scripts that integrate various external programs for processing ddRAD-seq data. The pipelines are available in https://github.com/CGRL-QB3-UCBerkeley/RAD. Paired-end raw fastq reads were first de-multiplexed based on the sequences of internal barcodes with a tolerance of one mismatch. The reads were then filtered to trim adapter contaminations and low-quality reads. The resulting cleaned reads were clustered with a sequence identity threshold of 0.95, and potential paralogs, loci containing repeats, and/or loci that were likely derived from incomplete restriction enzyme digestion were removed. The resulting RAD loci from each individual were then combined and collapsed into a non-redundant master reference set. Cleaned paired-end reads from each sample were aligned to the reference using Novoalign v.2 (http://www.novocraft.com/products/novoalign).

Phylogenetic inference was based on a maximum likelihood criterion. We used the GTRGAMMA nucleotide substitution mode with 1000 bootstrap replicates in RAxML v.8.1.16 (Stamatakis, 2014). The molecular dataset consists of a concatenated matrix with 1387 filtered loci and 7762

informative SNPs. Our analysis tested for the monophyly of *P. cordatum* and aimed at reconstructing phylogenetic relationships of *P. cordatum* with respect to *P. heptaphyllum s.l.*

Near-infrared spectroscopy (NIR). — NIR technology was used to test the hypothesis that *P. cordatum* represents a distinct entity from *P. heptaphyllum s.l.* For each specimen, a single spectrum was collected from three different dried leaflets using a desktop Spectroscopy Analyzer from Thermo Fisher Scientific, model Antaris II (Antaris, Waltham, Massachusetts, U.S.A.). A total of ten specimens of each taxon were included. All of these specimens were also included in the morphological and molecular phylogenetics analyses described above. Each spectrum represents the average of 16 scans including the absorbance of 1557 values sampled at intervals of 8 cm−1 within wavelengths of 4000–10,000 nm. An opaque black lid was placed over the reading area to avoid light scattering. A background calibration was performed automatically during every other reading. In total, 12 individuals of *P. cordatum* and 15 individuals of *P. heptaphyllum s.l.* were analyzed.

We used the Kolmogorov-Smirnov (KS) test to determine if the spectral curves of *P. cordatum* and *P. heptaphyllum s.l.* show significant differences. The KS-test is a non-parametric test that does not require any prior assumption about the distribution of the data (Lopes *et al.*, 2009). We ran the KS-test comparing all curves based on D-values within and across populations of *P. cordatum* and *P. heptaphyllum s.l.*

RESULTS

Morphological analyses. — The PCA based on vegetative and reproductive characters supports the hypothesis that *P. cordatum* is morphologically distinct from *P. heptaphyllum s.l.* The first two principal components were most useful for group discrimination (Fig. 2A). NMMs ignoring both principal components explained only 3% of the morphological variance. All NMMs indicated two distinct morphological groups based on the high BIC-values and the plot shows the highest empirical support (ordinate) and the optimum number of morphological groups (abscissa) supporting the hypothesis of two distinct species based on vegetative and reproductive characters. (Fig. 2B). Regarding vegetative traits, *P. heptaphyllum s.l.* has more pairs of leaflets (juga), longer petioles (Fig. 2C), higher specific leaf area (a proxy of leaf thickness), and higher plant height (Fig. 2D). *Protium cordatum* has shorter petals and lower flower density along the inflorescence axes (Fig. 2E) and a denser petal indumentum (Fig. 2F).

Phylogenomics. — *Protium cordatum* is strongly supported as monophyletic (BS = 100%; Fig. 3) and is sister to a clade that includes P. ovatum and P. dawsonii (BS = 100%) and the monophyletic *Protium heptaphyllum* species complex (BS = 100%). The sister-group relationship between *P. heptaphyllum s.l.* and *P. ovatum+P. dawsonii* is poorly supported (BS = 64%), meaning that there is uncertainty regarding the exact position of the clade *P. ovatum+P. dawsonii* with regard to *P. heptaphyllum* and *P. cordatum*. The population of *P. heptaphyllum* sampled in the neighboring forest across the habitat ecotone is distantly related to *P. cordatum* (samples from Fig. 1B are bolded in Fig. 3), indicating that both populations are genetically highly distinct despite the large potential for gene flow.

Leaf spectroscopy. — The KS-test showed significant (p-value < 2.2−16) dissimilarity between the spectra of *P. cordatum* and neighboring *P. heptaphyllum* populations (Fig. 4A), which

corresponds to the spectral discontinuity observed in the PCA ordination space (Fig. 4B). According to D-values, the spectral variability within populations of *P. cordatum* and *P. heptaphyllum s.l.* is significantly lower (p-value < 1.26−11) than across populations (Fig. 4C). Interestingly, the spectral region between 4000 and 5300 nm wavelength is more variable in terms of absorbance readings in comparison to 5300–7000 nm. In the latter, the variation within populations can be as high as among populations. Although the spectral signatures of *P. cordatum* and *P. heptaphyllum s.l.* are significantly different in the 7000–10,000 nm wavelength region, the overall absorbance variation within this spectral interval is lower than in the 4000–5300 nm region (coefficient of variation: 4000–5300 nm = 0.81 to 0.83; 7000–10,000 nm = 0.10 to 0.32).

DISCUSSION

In this study, we combined morphology, phylogenomics and spectroscopy to improve species delimitation within the *P. heptaphyllum* species complex. Data from different sources provide a consistent picture of the ideal taxonomic placement for *P. cordatum* within the Protieae tribe of Burseraceae (Fine *et al.*, 2014). Based on these multidisciplinary results, we conclude that *P. cordatum* should be reinstated as a formal species as initially described by Huber (1909), and not treated as an infraspecific taxon within *P. heptaphyllum*.

Taxonomic implications. — In 1909, *P. cordatum* was first described by Huber as a habitat specialist shrub in the white-sand savannas in Amazonia (also known as campinas). In 1992, D.C. Daly proposed a new status and a new combination of *P. cordatum* as a subspecies of *P. heptaphyllum*. He justified this decision by stating "there is a distinct geographic component to the differences between them, but they can be distinguished only by the rather quantitative characters" and concluded that "further material of subsp. *cordatum* is needed before their differences can be defined adequately and the transfer made" (Daly, 1992: 298).

We analyzed additional material of *P. cordatum*, including additional samples from the type locality, and present a key with the most relevant differences of discrete and continuous morphological traits among *P. cordatum* and *P. heptaphyllum s.l.* We also found that vegetative (leaf petiole, number of juga, and specific leaf area) and reproductive characters (flower density, petal indumenta, corolla length) are discontinuous with minimal phenotypic overlap. Our phylogenetic results showed that *P. cordatum* is not closely related to *P. heptaphyllum*, and thus is likely genetically distinct from populations of *P. heptaphyllum* sampled at parapatric habitat ecotones. Individuals of *P. cordatum* sampled from the exact type locality represent a monophyletic clade to the exclusion of all other lineages of *P. heptaphyllum s.l.*, and the *P. ovatum*+*P. dawsonii* clade.

We found that *P. cordatum* was not closely related to the other white-sand specialist taxon within the *P. heptaphyllum* complex, *P. heptaphyllum* subsp. *ulei*, which presents consistent morphological differences (i.e., habit, leaf shape, floral and fruit traits) and a broader geographic distribution over the Amazon basin (Peru, Venezuela, Guyana, Central and Northern Amazonia). A study based on gene flow estimates and hybridization tests including a larger sample size was conducted to investigate the species limits within the *P. heptaphyllum* clade, including subsp. *ulei*, and an updated taxonomic treatment is in preparation as a monograph (Damasco *et al.*, in prep.). Here, we focus solely on the reestablishment of the *P. cordatum* lineage at species rank due to the

strong support for it being morphologically distinct and outside the clade corresponding to *P. heptaphyllum s.l.*

The importance of in-depth integrative studies of plant lineages. — Our results directly address concerns regarding the "hyperdominance phenomenon" in the Neotropics (Ter Steege *et al.*, 2013; Cardoso *et al.*, 2017). Based on the plot-dataset published by the Amazon Tree Diversity Network, *P. heptaphyllum s.l.* is classified as the eleventh most dominant taxon in the Amazon, especially common in the white-sands in the upper Rio Negro basin and the Guiana Shield. But, as we demonstrate here, a taxon that was considered part of a hyperdominant clade represents at least two independent lineages. If we ignore the possibility that dominant clades may include different putative species, we could seriously underestimate the diversity of tree species and make errors predicting the relative abundances of plant communities in the Neotropics.

Although there has been a great effort to estimate the accurate diversity of plants in the Amazon and the Americas (Cardoso *et al.*, 2017; Ulloa *et al.*, 2017), as many as 10%–20% of species may remain undescribed in the tropics (Pimm *and* Joppa, 2015). While it is possible that the majority of these unknown species are located in areas that have yet to be visited by botanists, we suggest that the number of undescribed species is likely to be even higher if we consider that dominant and widespread plant species, like *P. heptaphyllum s.l.*, may contain many hidden lineages that should not be considered conspecific. Our findings highlight the importance of additional in-depth studies of individual Neotropical species, as well as the importance of using multiple lines of evidence in taxonomy to delimit taxa which will, in turn, lead to more accurate estimates of species diversity in the tropics.

Implications for the near-infrared (NIR) technology. — Near-infrared technology has been used by plant taxonomists to discriminate several Neotropical plant groups. Recent studies of two diverse Amazonian genera, Eschweilera (Durgante *et al.*, 2013) and *Protium* (Lang *et al.*, 2015), showed that NIR spectra could accurately discriminate distinct species with over 96% success. Regarding closely related plant groups, this technology has been effective to discriminate cryptic species within Pagamea (Rubiaceae) (Prata *et al.*, 2018). In addition to the morphological and phylogenetic results, the NIR data indicate that *P. cordatum* and *P. heptaphyllum* have distinct spectral signatures. Even though they are closely related lineages, the spectral variation within each taxon was significantly lower than the spectral variation between taxa.

The leaf spectra can be correlated with the chemical composition and the anatomical structure inside the leaves (Asner *and* Martin, 2008; Féret *and* Asner, 2013). More specifically, changes in cell wall composition, such as polysaccharides, proteins, and phenolic compounds are believed to be evolutionarily conserved among different species (Asner *et al.*, 2014). As NIR application to tropical botany advances, more research is needed to understand better the characteristics behind the spectral values and which factors might have a significant impact on leaf absorbance signatures along the spectrum. For instance, we found little variation across a large section of the wavelength spectrum, as well as high redundancy of absorbance values among similar spectral regions or neighboring wavelength sites (as also noticed by Durgante *et al.*, 2013). More studies are needed to optimize the usage of spectral data in plant taxonomy, and we believe that future research using NIR technology should investigate what the most informative regions of the wavelength spectrum are and whether there is an optimum set of wavelength bands that may increase the accuracy of discriminant models in taxonomy.

TAXONOMY

Protium cordatum Huber in Bol. Mus. Goeldi Hist. Nat. Ethnogr. 5: 433. 1909 ≡ *P. heptaphyllum* subsp. *cordatum* (Huber) Daly in Brittonia 44: 298. 1992 – Lectotype (designated by Swart in Recueil Trav. Bot. Néerl. 39: 330. 1942): Brazil. Pará: Faro, Campo do Tigre, 21 Aug 1907 (m fl), Ducke s.n. (B [destroyed]; isolectotypes: F [photo!], MG No. 8463!, NY barcode 00345733!, RB No. 20522!).

Description. – Shrubs ca. 1.5–3 m tall, crown open. Stems highly branched from base; outer bark light to dark gray, thin, often rough from high density of lenticels and light-colored lichens, inner-bark white or light yellow; branchlets, striate and brown towards apices, lenticels sparse. Resin flammable, transparent, and viscous when fresh, dark grey with crystalline texture when dry. Leaves glabrous, ca. 6–12 cm long, often 1–3-jugate; petiole ca. 2–7 cm long, 1–2 mm diam. near base, often striate with appressed fine hairs to 0.05 mm long; interjuga 1–1.5 cm long; basal petiolules 3–9 mm long, other lateral petiolules 2–6.5 mm long, terminal petiolule 4–15 mm long; lateral and distal pulvinuli inconspicuous; leaflet blades ca. 5–9.5 cm long, 1.5–5 cm wide, elliptic to ovate, highly coriaceous, drying dark green to reddish brown abaxially, grey to green or light brown adaxially, faces dull, apex cuspidate or rarely acute, the acumen to 10 mm long, base cordate or occasionally rounded, often asymmetric, margin entire, secondary vein framework festooned-brochidodromous, costal secondaries in 6–14 pairs, the spacing irregular, decreasing toward apex and base, the angle slightly decreasing toward base, course essentially straight, occasionally one intersecondary vein per pair of costal secondaries and parallel to them, adaxial face with midvein narrowly prominulous, secondary veins mostly flat, tertiaries flat, irregular-polygonal, quaternaries flat, irregular-reticulate, abaxial face with midvein and secondaries prominulous, tertiaries mostly prominulous, irregular-polygonal, quaternaries flat, irregular-reticulate.

Staminate inflorescences, 4–15 mm long, 4–10 mm diam. near base, secondary axes 2–9 mm long, all axes with dense malpighiaceous hairs to 0.1 mm long, bristles (short, fine, erect white hairs) also present; bracts 0.3–0.7 mm long on primary axes and 0.2–0.6 mm on secondary axes, elliptic to deltate, apex acute; bracteoles 0.1–0.3 mm, coriaceous, with dense, thick, white malpighiaceous hairs; pedicel 0.5–1.5 mm long, 0.2–0.6 mm diam., cylindrical, with pubescence as on inflorescence axes. Staminate flowers 4-merous, 2–3 mm long; calyx 0.4–0.6 mm long, 1.5–2.5 mm diam., exceeding disk or nearly equal, not divided to base, the lobes mostly inconspicuous and separated by a flat sinus, few flowers with visible lobes 0.2–0.5 mm long, 0.7–1 mm wide with occasional acute to acuminate apex, abaxial pubescence as on inflorescence axes; corolla ovate to urceolate, 1.3–2.2 mm long, 0.5–1 mm wide, mainly light yellow with occasional orange-reddish tonality (especially in bud), apiculum 0.1–0.3 mm long, broadly ovate, somewhat coriaceous, inflexed, abaxial pubescence as on calyx except trichomes longer and denser toward apex, adaxially mostly glabrous or with sparse bristles, margin sparsely papillate; stamens 8, equal, inserted on outer edge of disk, 1–1.8 mm long, anthers 0.4–0.6 mm long, oblong-ovate in dorsiventral view, lanceolate in profile, filaments cylindrical to compressed with dense bristles; annular disk globose, 0.5–0.9 tall, glabrous or with sparse bristles, essentially discoid with narrowly conical center; pistillode 0.2–0.4 mm, exceeding disk; pistillode with high density of long malpighiaceous hairs ca. 0.3 mm (longer relative to corolla indumentum).

Pistillate inflorescences 5–15 mm long, 5–8 mm diam. near base, secondary axes 4–9 cm long; bracts 0.4–0.7 mm long on primary axes and ca. 0.4 mm long on secondary axes, ovate to rarely

deltate, apex acute to acuminate; bracteoles 0.1–0.3 mm, coriaceous with dense, thick white malpighiaceous hairs; pedicel 0.5–1.5 mm long, 0.2–0.6 mm diam., cylindrical, with pubescence as on inflorescence axes. Pistillate flowers 2.1–2.8 mm long; calyx 0.3–0.6 mm, 1.1–1.8 mm diam., height relative to disk as on staminate flowers, pubescence as on inflorescence axes; corolla 1.5–2 mm long, 0.8–1.2 mm wide, ovate to urceolate, somewhat coriaceous, apiculum 0.2–0.3 mm long; staminode insertion and shape as on staminate flowers, 0.8–1.2 mm long, anthers 0.4–0.5 mm long; annular disk 1–1.2 mm long, 0.9–1.1 mm diam., glabrous or with sparse malpighiaceous hairs; style 1.1–1.5 mm long, stigma 0.2–0.4 mm long, sessile, erect, depressed-globose; ovary globose-ovoid, glabrous or with sparse malpighiaceous hairs.

Fruit maturing red, mostly globose to slightly oblique-ovoid, ventricose, dry size ca. 6–8 mm long, 4–6 mm diam. (1 locule), 8–10 mm long, 7–9 mm diam. (2–4 locules), glabrous with sparse bristle hairs near the base, smooth, drying slightly wrinkled, apex mostly obtuse to round (in globose fruits), occasionally acute (in ovoid fruits), base slightly substipitate (stipe ca. 0.8–1 mm long), rounded to truncate above stipe. Fruiting pedicel 0.5–1.5 mm long, 0.4–0.6 mm diam., cylindrical.

Distribution and habitat. – *Protium cordatum* is a rare shrub endemic to white-sand savannas and sandy riverbanks. This species is most common in seasonally flooded areas of white-sand ecosystems (G. Damasco, pers. obs.). It occurs in white-sand areas of the Nhamundá River (Fig. 1C), Faro municipality, and was previously reported from the white-sand savannas near the Tocantins River, Tucuruí municipality, Pará, Brazil. We did visit the Tucuruí region and looked for populations of *P. cordatum* in areas where the specimens have been collected before but we could not find them. We examined the specimens collected in Tucuruí, and they were morphologically identical to populations collected in Faro (the type locality). There is a good chance that additional populations of *P. cordatum* could be found in white-sand patches located nearby Oriximiná or savannas near Santarém (both in Pará State, Brazil).

Uses. – No uses are reported, but *P. cordatum* contains copious amounts of resin like many species of Burseraceae. In many species of *Protium*, these resins are often burned as light sources, incense or used as medicines (e.g., Siani *et al.*, 2012).

Diagnostic features. – *Protium cordatum* is morphologically similar to *P. heptaphyllum s.l.* but differs with its shrubby habit (vs. stunted to tall tree habit in *P. heptaphyllum s.l.*) and coriaceous leaflets that are disposed perpendicular to the leaf branch axis (vs. chartaceous and often smooth leaflets with a non-perpendicular disposition in *P. heptaphyllum s.l.*) (Fig. 5M). Furthermore, the leaf petiole is shorter in *P. cordatum* (77.4 ± 7.3 mm) than in *P. heptaphyllum s.l.* (110.8 ± 19.3 mm), as is the terminal petiolule (26.9 ± 3.5 mm and 53.8 ± 15 mm, in *P. cordatum* and *P. heptaphyllum s.l.* respectively). The secondary veins of *P. cordatum* are usually impressed rather than prominent in *P. heptaphyllum s.l.* and tertiary venation is barely visible in *P. cordatum* due to darker coloration and flatness at the adaxial face compared to the abaxial face (Fig. 5K,L). The inflorescence has fewer flowers (4–8) than *P. heptaphyllum s.l.* (6–16) and the corolla length is usually shorter in *P. cordatum* (up to 15 mm) than in *P. heptaphyllum s.l.* (up to ca. 70 mm). The calyx and corolla of *P. cordatum* have a high density of malpighiaceous hairs on the abaxial surfaces, while the calyx and corolla of *P. heptaphyllum s.l.* are glabrous or sparsely bristly. In addition, the pistillode in the staminate flowers of *P. cordatum* has long malpighiaceous hairs while the pistillode is glabrous or sparsely bristly in *P. heptaphyllum s.l.* Fruits of *P. cordatum* are often globose, whereas *P. heptaphyllum s.l.* often has obliquely ovoid fruit with an occasionally

obtuse apex. The resin is more viscous and darker when dry in *P. cordatum* in comparison to that of *P. heptaphyllum s.l.*, which is more transparent and more abundant and waterier. Diagnostic features for *P. ovatum* and *P. dawsonii* will be covered in a future publication because more samples are needed to review their taxonomy. Both taxa are savanna specialists inhabiting nutrient-scarce soils in the Brazilian Cerrado (habitat also known as cerrado sensu stricto). *Protium ovatum* and *P. dawsonii* are found as stunted shrubs, and their leaflets are usually ovate and often have a serrate margin. According to Jose Cuatrecasas (author that described *P. dawsonii*), *P. ovatum* differs from *P. dawsonii* by having hairs present on the adaxial leaflet surface. However, after examining several herbarium specimens in NY, this morphological character is not consistent in *P. dawsonii*. A detailed taxonomic revision including more samples is necessary to resolve the taxonomy of the clade *P. ovatum*+*P. dawsonii*.

Additional specimens examined. – Brazil. Pará, Campina de Santa Rosa, lat −3.7661, long −49.6725, J. Ramos 626 (INPA); Campina de Santa Rosa, ramal da BR-422, lat −3.7661, long −49.6725, J. Ramos 1147 (INPA); Campinas de Santa Rosa, lat −3.7661, long −49.6725, J. Revilla 8502 (INPA); Campos a Leste de Faro, A. Ducke s.n. (MG); Margem direita da BR-263, km 16, lat −3.75105, long −49.5473, M.G. Silva 5503 (INPA, MG); BR-263, km 16, lat −3.75105, long −49.5473, M.G. Silva 5806 (INPA, MG); Approx. 25 km S of Tucuruí, just off the old BR-422 at the junction with an abandoned railroad bed, lat −3.99065, long −49.6736, D.C. Daly 1080 (INPA, NY); Margem direita da PA-149, km 35, lat −3.76457, long −49.6736, J. Ramos 883 (INPA); PA-149, lat −3.76457, long −49.6736, F.E.L. Miranda 395 (INPA); PA-149, lat −3.76457, long −49.6736, F.E.L. Miranda 408 (INPA).

Nomenclatural notes. – Huber described *Protium cordatum* in 1909 with a citation of only one specimen, indicated as "Ducke 8463", 21 Aug 1907, but with no mention of the type or the herbarium in which the specimen was deposited. Later in 1942, in a review of *Protium* and allied genera in Burseraceae, J.J. Swart included two specimens under the name *P. cordatum*, one of which is the specimen ("Ducke 8463") cited by Huber (1909). In his review, Swart stated that both specimens were deposited in B. In 1992, upon making the new combination *P. heptaphyllum* subsp. *cordatum*, Daly indicated that the number "8463" of Ducke's collection cited by Huber (1909) was actually the catalog number of the MG Herbarium, where Huber was working at the time, and that the correct collection number of this specimen should be "Ducke 20522". Daly also noted three isolectotypes deposited in MG, NY, and RB and declared that the lectotype designated by Swart deposited in B was almost certainly destroyed. After a new careful inspection of all isolectotypes mentioned in Daly (1992), we conclude that there is no collection number associated with Ducke's specimen. The numbers "8463" and "20522" correspond to specimen catalog numbers at MG and RB, respectively. In addition, both Swart (1942) and Daly (1992) mentioned in the protologue that the publication year of *P. cordatum* was 1908. Volume 5 of the Boletim do Museu Goeldi de Historia Natural e Ethnographia was organized in two fascicles, the first one dated February 1908, and the second one March 1909. After examining the original manuscript, we noticed that the description of *P. cordatum* was published in the second fascicle. Therefore, the correct year of Huber's publication is 1909, rather than 1908.

Key to the identification of *Protium cordatum* and members of the *P. heptaphyllum* species complex

1. Shrubs ca. 1.5–3 m tall, terminal petiolules 4–15 mm, lateral and terminal pulvinuli inconspicuous; leaflet blades often coriaceous, base cordate, rarely truncate and rounded; tertiary veins flat on adaxial surface; petals to 2.5 mm long with dense malpighiaceous hairs towards the apex; anthers 0.45–0.6 mm; pistillode with malpighiaceous hairs to 0.3 mm; fruit to 1.5 cm ……..... ***Protium cordatum***

1. Treelets and trees to 15 m tall; terminal petiolules 5–45 mm, lateral and terminal pulvinuli often present; leaflet blades chartaceous to moderately coriaceous, base acute to obtuse; tertiary veins prominent to prominulous on adaxial surface; petals to 4.2 mm long with glabrous or with scattered appressed hairs; anthers 0.45–1.2 mm; pistillode glabrous or less often with scattered ascending hairs; fruit to 2.3 cm …….. **2**

2. Leaflets drying greenish-tan or brown, chartaceous, margin flat and entire near base; leaflet tertiary veins prominulous on abaxial surface; petals lanceolate, 2.8–4.2 mm; anthers 0.7–1.2 mm; pistil 1.7–2.5 mm, style 0.8–1.1 mm; fruit 1.4–2.3 …..... ***P. heptaphyllum*** subsp. ***heptaphyllum***

2. Leaflets drying reddish with slightly caudate to acuminate apex, moderately coriaceous, margin often revolute near base; leaflet tertiary veins prominulous to flat on abaxial surface; petals lanceolate to ovate, 2–2.6 mm; anthers 0.45–0.6 mm; pistil 1.2–1.7 mm, style 0.5–0.75 mm; fruit 1–1.4 cm ..... ***P. heptaphyllum*** subsp. ***ulei***

REFERENCES

Asner, G.P., and Martin, R.E. 2008. Spectral and chemical analysis of tropical forests: Scaling from leaf to canopy levels. Remote Sensing Environm. 112: 3958–3970.

Asner, G.P., Martin, R.E., Carranza-Jiménez, L., Sinca, F., Tupayachi, R., Anderson, C.B., and Martinez, P. 2014. Functional and biological diversity of foliar spectra in tree canopies throughout the Andes to Amazon region. New Phytol. 204: 127–139.

Bik, H.M. 2017. Let's rise up to unite taxonomy and technology. PLOS Biol. 15: e2002231.

Cadena, C.D., Zapata, F., and Jiménez, I. 2018. Issues and perspectives in species delimitation using phenotypic data: Atlantean evolution in Darwin's finches. Syst. Biol. 67: 181–194.

Cardoso, D., Särkinen, T., Alexander, S., Amorim, A.M., Bittrich, V., Celis, M., Daly, D.C., Fiaschi, P., Funk, V.A., and Giacomin, L.L. 2017. Amazon plant diversity revealed by a taxonomically verified species list. Proc. Natl. Acad. Sci. U.S.A. 114: 10695–10700.

Daly, D.C. 1992. New taxa and combinations in Protium Burm. f. Studies in neotropical Burseraceae VI. Brittonia 44: 280–299.

Dayrat, B. 2005. Towards integrative taxonomy. Biol. J. Linn. Soc. 85: 407–415.

Durgante, F.M., Higuchi, N., Almeida, A., and Vicentini, A. 2013. Species spectral signature: Discriminating closely related plant species in the Amazon with near-infrared leaf-spectroscopy. Forest Ecol. Managem. 291: 240–248.

Féret, J.-B., and Asner, G.P. 2013. Tree species discrimination in tropical forests using airborne imaging spectroscopy. IEEE Trans. Geosci. Remote Sensing 51: 73–84.

Fine, P.V.A., Zapata, F., Daly, D.C., Mesones, I., Misiewicz, T.M., Cooper, H.F., and Barbosa, C. 2013. The importance of environmental heterogeneity and spatial distance in generating phylogeographic structure in edaphic specialist and generalist tree species of Protium (Burseraceae) across the Amazon basin. J. Biogeogr. 40: 646–661.

Fine, P.V.A., Zapata, F., and Daly, D.C. 2014. Investigating processes of neotropical rain forest tree diversification by examining the evolution and historical biogeography of the Protieae (Burseraceae). Evolution 68: 1988–2004.

Fišer, C., Robinson, C.T., and Malard, F. 2018. Cryptic species as a window into the paradigm shift of the species concept. Molec. Ecol. 27: 613–635.

Gonzalez, M.A., Baraloto, C., Engel, J., Mori, S.A., Pétronelli, P., Riéra, B., Roger, A., Thébaud, C., and Chave, J. 2009. Identification of Amazonian trees with DNA barcodes. PLOS ONE 4: e7483.

Huber, J. 1909. Materiaes para a Flora amazonica VII. Plantae Duckeanae austro-guyanenses. Bol. Mus. Goeldi Hist. Nat. Ethnogr. 5(2): 294–436.

Kress, W.J., Wurdack, K.J., Zimmer, E.A., Weigt, L.A., and Janzen, D.H. 2005. Use of DNA barcodes to identify flowering plants. Proc. Natl. Acad. Sci. U.S.A. 102: 8369–8374.

Lang, C., Costa, F.R.C., Camargo, J.L.C., Durgante, F.M., and Vicentini, A. 2015. Near Infrared Spectroscopy facilitates rapid identification of both young and mature Amazonian tree species. PLOS ONE 10: e0134521.

Lang, C., Almeida, D.R., and Costa, F.R. 2017. Discrimination of taxonomic identity at species, genus and family levels using Fourier Transformed Near-Infrared Spectroscopy (FT-NIR). Forest Ecol. Managem. 406: 219–227.

Leaché, A.D., Fujita, M.K., Minin, V.N., and Bouckaert, R.R. 2014. Species delimitation using genome-wide SNP data. Syst. Biol. 63: 534–542.

Lopes, R.H., Reid, I., and Hobson, P.R. 2009. The two-dimensional Kolmogorov-Smirnov test. In: XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research, April 23–27 2007, Amsterdam, the Netherlands. PoS(ACAT) 045.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., and Hoekstra, H.E. 2012. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLOS ONE 7: e37135.

Pimm, *S.L.*, and Joppa, L.N. 2015. How many plant species are there, where are they, and at what rate are they going extinct? Ann. Missouri Bot. Gard. 100: 170–176.

Prata, E.M.B., Sass, C., Rodrigues, D.P., Domingos, F.M.C.B., Specht, C.D., Damasco, G., Ribas, C.C., Fine, P.V.A., and Vicentini, A. 2018. Towards integrative taxonomy in Neotropical botany: Disentangling the Pagamea guianensis species complex (Rubiaceae). Bot. J. Linn. Soc. 188: 213–231.

Schlick-Steiner, B.C., Steiner, F.M., Seifert, B., Stauffer, C., Christian, E., and Crozier, R.H. 2010. Integrative taxonomy: A multisource approach to exploring biodiversity. Annual Rev. Entomol. 55: 421–438.

Schwarz, G. 1978. Estimating the dimension of a model. Ann. Statist. 6: 461–464.

Scrucca, L., and Raftery, A.E. 2014. clustvarsel: A package implementing variable selection for model-based clustering in R. arXiv: 1411.0606.

Scrucca, L., Fop, M., Murphy, T.B., and Raftery, A.E. 2016. mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. R Journal 8: 289–317.

Siani, A., Nakamura, M.J., Tappin, M., Monteiro, S., Guimarães, A., and Ramos, M. 2012.

Chemical composition of South American Burseraceae non-volatile oleoresins and preliminary solubility assessment of their commercial blend. Phytochem. Analysis 23: 529–539.

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

Swart, J.J. 1942. A monograph of the genus Protium and some allied genera (Burseraceae). Recueil Trav. Bot. Néerl. 39: 211–446.

Ter Steege, H., Pitman, N.C., Sabatier, D., Baraloto, C., Salomão, R.P., Guevara, J.E., Phillips, O.L., Castilho, C.V., Magnusson, W.E., Molino, J., Monteaugo, A., Vargas, P.N., Montero, J.C., Feldpausch, T.R., Coronado, E.N.H., Kileen, T.J., Mostacedo, B., Vasquez, R., Assis, R.L., Terborgh, J., Wittmann, F., Andrade, A., Laurance, W.F., Laurance, S.G.W., Marimon, B.S., Marimon, B., Jr., Vieira, I.C.G., Amaral, I.L., Brienen, R., Castellanos, H., López, D.C., Duivenvoorden, J.F., Mongollón, H.F., Matos, F.D.A., Dávila, N., García-Villacorta, R., Diaz, P.R.S., Costa, F., Emilio, T., Levis, C., Schietti, J., Souza, P., Alonso, A., Dallmeier, F., Montoya, A.J.D., Piedade, M.T.F., Araujo-Murakami, A., Arroyo, L., Gribel, R., Fine, P.V.A., Peres, C.A., Toledo, M., Aymard, G.A., Baker, T.R., Cerón, C., Engel, J., Henkel, T.W., Maas, P., Petronelli, P., Stropp, J., Zartman, C.E., Daly, D., Neil, D., Silveira, M., Paredes, M.R., Chave, J., Filho, D.A.L., Jørgensen, P.M., Fuentes, A., Schöngart, J., Valverde, F.C., Di Fiore, A., Jimenez, E.M., Mora, M.C.P., Phillips, J.F., Rivas, G., Van Andel, T.R., von Hildebrand, P., Hoffman, B., Zent, E.L., Mahli, Y., Prieto, A., Rudas, A., Ruschell, A.R., Silva, N., Vos, V., Zent, S., Oliveira, A.A., Schutz, A.C., Gonzales, T., Nascimento, M.T., Ramirez-Angulo, H., Sierra, H., Tirado, M., Medina, M.N.U., Van der Heijden, G., Vela, C.I.A., Torre, E.V., Vriesendorp, C., Wang, O., Young, K.R., Baider, C., Baslev, H., Ferreira, C., Mesones, I., Torres-Lezana, A., Giraldo, L.E.U., Zagt, R., Alexiades, M.N., Hernandez, L., Huamantupa-Chuquimaco, I., Milliken, W., Cuenca, W.P., Pauletto, D., Sandoval, E.V., Gamarra, L.V., Dexter, K.G., Feeley, K., Lopez-Gonzales, G., and Silman, M.R. 2013. Hyperdominance in the Amazonian tree flora. Science 342: 1243092.

Ulloa, C., Acevedo-Rodríguez, P., Beck, S., Belgrano, M.J., Bernal, R., Berry, P.E., Brako, L., Celis, M., Davidse, G., Forzza, R.C., Gradstein, S.R., Hokche, O., León, B., León-Yánez, S., Magill, R.E., Neill, D.A., Nee, M., Raven, P.H., Stimmel, H., Strong, M.T., Villaseñor, J.L., Zarucchi, J.L., Zuloaga, F.O., and Jørgensen, P.M. 2017. An integrated assessment of the vascular plant species of the Americas. Science 358: 1614–1617.

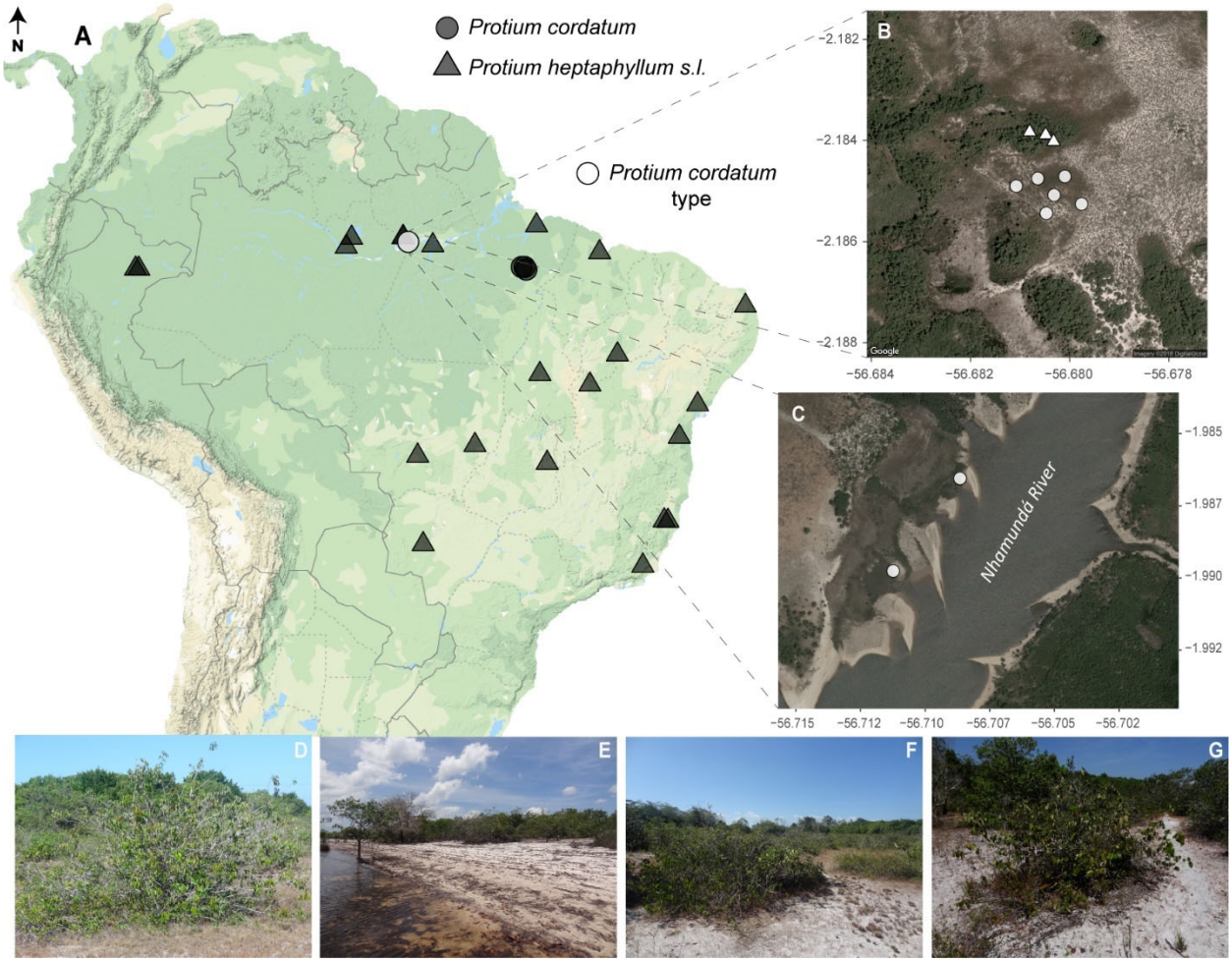

Figure 1. A) Distribution of the specimens included in this study. Circles represent *Protium cordatum* (white circle: coordinates of *P. cordatum* from Tucuruí, Pará, Brazil, not sampled in this study) and triangles represent *Protium heptaphyllum sensu lato*. B) Populations sampled in this study along a parapatric habitat ecotone (white circle: *Protium cordatum* individuals; white triangle: *Protium heptaphyllum sensu lato* individuals) at the exact location where the lectotype was collected. C) Individuals of *Protium cordatum* sampled at the margin of the Nhamunda River. D-G) Examples of habitat and photos of *Protium cordatum*, a native and endemic shrub of Amazonian white-sand savannas.

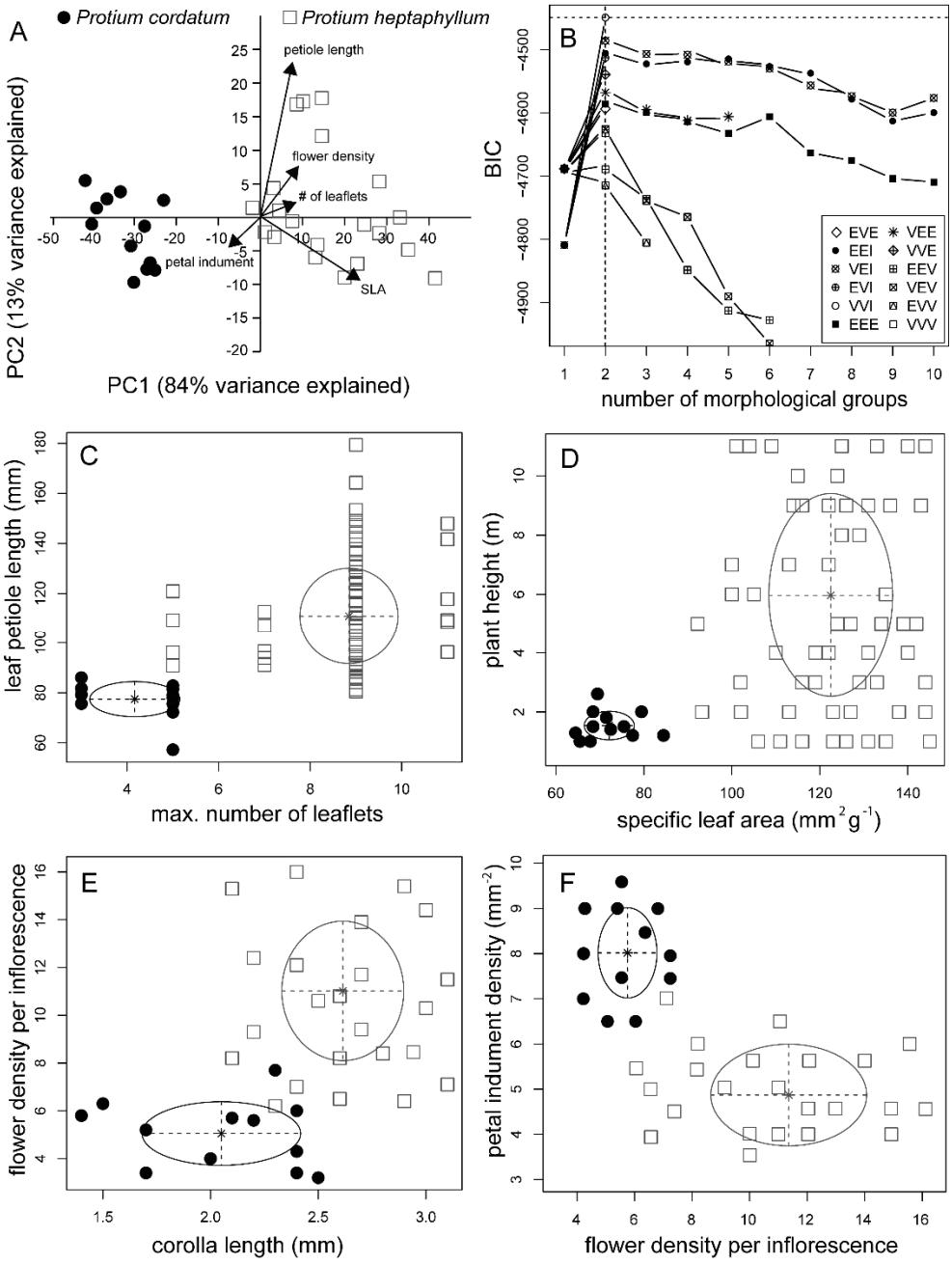

Figure 2. A) Principal Component Analysis (PCA) of morphological characters of *Protium cordatum* and *Protium heptaphyllum sensu lato*. Arrows correspond to PCA loadings of most informative morphological characters. B) BIC from normal mixture model (NMM) analysis using 12 model parameterizations and up to 10 morphological groups. Different symbols and line types encode different model parameterizations. C to F) A projection of the morphological characters, with different symbols indicating the classification corresponding to the best model as determined by the NMM analysis. The component means are marked, and ellipses with axes are drawn corresponding to their covariances. Vegetative traits were measured on 56 individuals of *P.heptaphyllum sensu lato* and 12 individuals of *P. cordatum*. Reproductive traits were measured on 23 individuals of *P. heptaphyllum* and 12 individuals of *P. cordatum*.

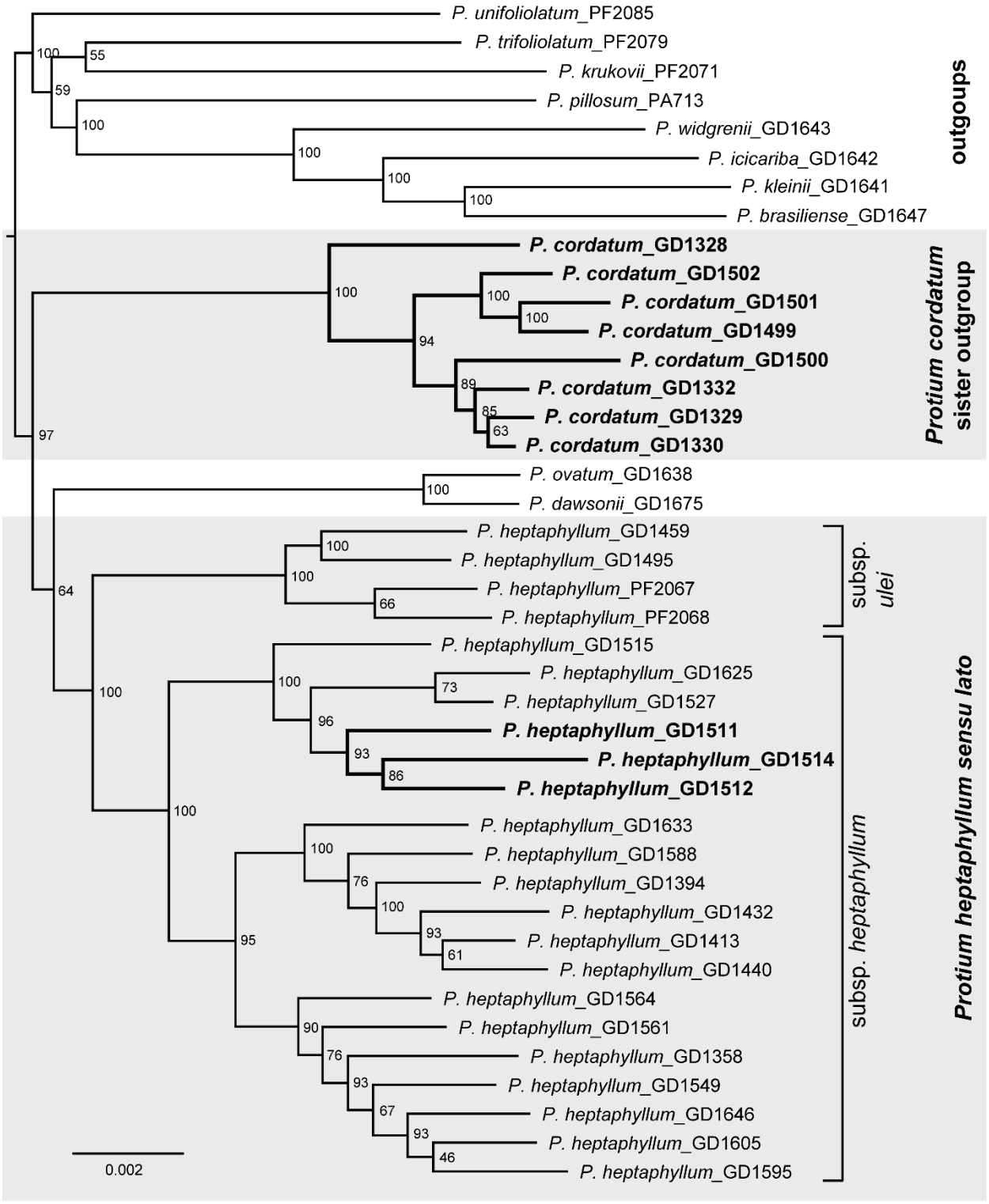

Figure 3. Maximum likelihood phylogeny based on a genome-wide ddRAD-seq dataset with 1,333 filtered loci and 21,359 informative SNPs. Grey boxes highlight the monophyletic clade of *Protium cordatum* as a new outgroup of *Protium heptaphyllum sensu lato*. Bolded branches in the phylogeny correspond to individuals sampled in the parapatric habitat ecotone shown in Figure 1B.

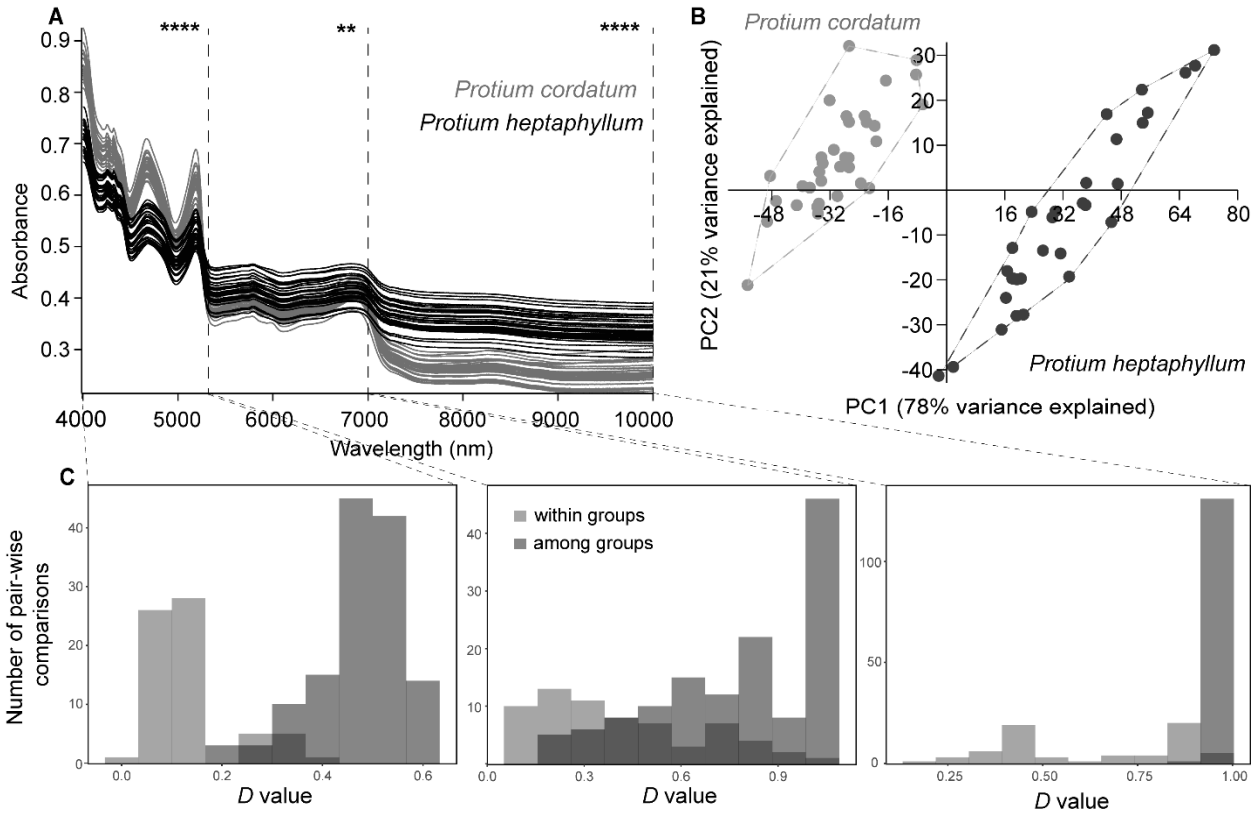

Figure 4. A) Representation of full near infra-red (FT-NIR) spectra for ten specimens of *Protium cordatum and Protium heptaphyllum sensu lato*, respectively. The spectral readings are expressed as absorbance values between the wavelength 4000 to 10,000 nm and each spectrum consists of an average of 1557 absorbance values for three different leaflets per specimen. The full spectra were subdivided into three spectral regions due to different absorbance variation within and across populations. B) Two-dimensional Principal Component Analysis based on FT-NIR spectral data. C) Histograms of pair-wise comparisons indicating the absorbance similarity based on D-values estimates within and across populations and compared among three spectral regions (4,000-5,300 nm; 5,400-7,000 nm; 7,000-10,000 nm).

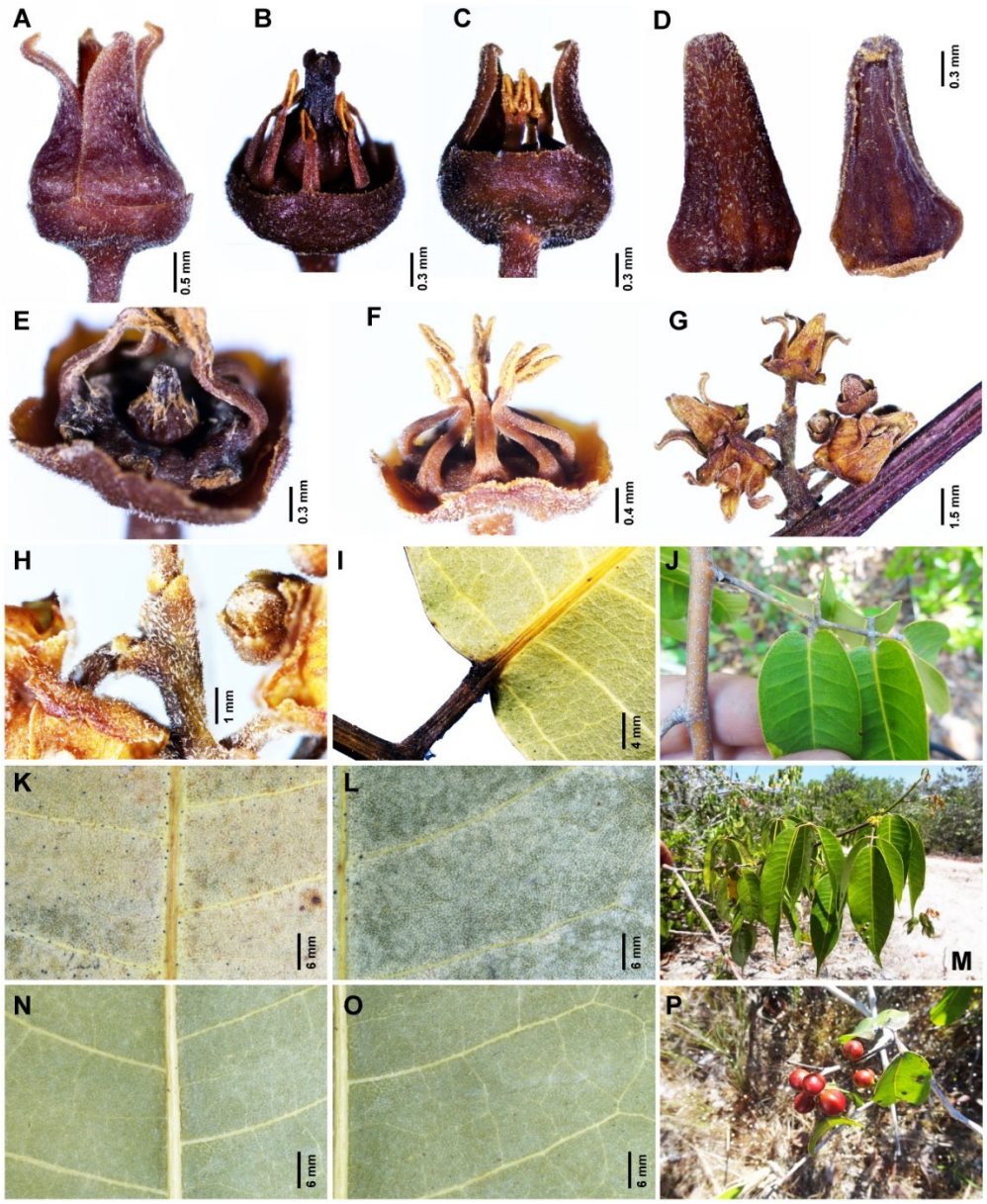

Figure 5. A) Flowers 4-merous with corolla molding a urceolate shape. B) Pistillate flowers with globose ovary and stigma erect and sessile. C) Staminate flower with calys exceeding disk or nearly equal, not divided to base, the lobes mostly inconspicuous and separated by a flat sinus. D) Corolla with dense, thick, white malpighiaceous hairs abaxially and mostly glabrous or with sparse bristle hairs adaxially. E) Pistillode with high density of long malpighiaceous hairs. F) Anthers lanceolate in profile and filaments cylindrical to compressed with dense bristles. G) Inflorescence with all bracts and bracteoles coriaceous and dense pubescence as on the corolla. H) Bracteoles coriaceous with dense, thick, white malpighiaceous hairs. I) Dry leaflet with cordate base or occasionally rounded and often asymmetric. J) Fresh leaflets collected in the field with cordate base. K and L) The secondary veins are usually impressed rather than prominent and tertiary venation is barely visible due to darker coloration and flatness at the adaxial face compared to the abaxial face (N and O). M) Fresh stem and leaves with 3 juga collected in the field. N and O) Leaflet abaxial face. P) Mature fruits mostly globose to slightly oblique-ovoid.

CHAPTER 3 Leaf transcriptome assembly of *Protium copal* (Burseraceae) and annotation of terpene biosynthetic genes

Abstract Burseraceae plant species are globally recognized for producing resins and essential oils with medical properties and economic values. In addition, most of the aromatic and non-aromatic properties of Burseraceae resins are composed majorly by terpene and terpenoid chemicals. Although terpene genes have been identified in model plant crops (e.g. *Citrus*, *Arabidopsis*), very few genomic resources are available for non-model groups, including the highly-diverse Burseraceae family. Here, we aim to assemble a leaf transcriptome of *Protium copal*, a widespread aromatic tree in Central America, and describe the functional annotation of putative terpene biosynthetic genes. The genomic resources of *Protium copal* can potentially be used to generate novel sequencing markers applied to population genetics and comparative phylogenetics studies in Burseraceae and to investigate the diversity and evolution of terpene chemicals in tropical plant species. In addition, the identification of terpene genes in *Protium copal* are relevant for understanding the synthesis of economically important terpene products in Burseraceae.

Keywords RNA-sequencing; de novo assembly; *Protium*; terpenoid; copal resin.

INTRODUCTION

The Burseraceae family harbors over 600 species of plants which produce resins and essential oils that have been economically and medicinally important for millennia in the Neotropics as well as parts of Asia and Africa. Uses of scents such as myrrh (*Commiphora* spp.) and frankincense (*Boswellia* spp.) are mentioned in biblical texts and are still used today in many religious rituals (Langenheim *et al.* 2003). South and Central America Mayan records dating back to 600 BC describe the use of copal (*Protium* sp. and *Bursera* sp.) resins as incense and medicines (Stacey *et al.* 2006). Today, Burseraceae plant species continue to be recognized globally for their medical, aromatic, flammable, and adhesive properties.

The majority of resins and oils produced by Burseraceae plants are terpenes scents. Most of these are produced internally by plant secretory cells and are defined as aromatic soluble mixtures of volatiles and non-volatiles rich in terpenoids. While the production of hundreds of different terpene chemicals is conserved among all plants, different lineages are also known to produce their own specialized terpenes. Currently, described terpenes already exceed 60,000 (Cheng *et al.* 2007) and as new specialized terpenes are discovered and described, that number will continue to increase (Chen *et al.* 2011). In highly diverse plant families such as Burseraceae, terpenes include monoterpenes, sequiterpenes, diterpenes, and triterpenes found in the shoots, leaves, and flowers (Langenheim *et al.* 2003).

Terpene chemicals play major biological roles in plant metabolism and life cycle. For instance, primary terpenoids are major constituents of plant membranes and are important for maintaining the fluidity of these cells (Pichersky *et al.* 2018). Defense against natural enemies (i.e. herbivores, pathogens, viruses) is another key function of terpenoids, whether directly targeting enemies as toxins and repellents, or indirectly through the attraction of predators or parasitoids of such enemies (Kessler *et al.* 2011). Although the functional aspects of terpene chemicals as the main

defense strategy of plants are well known, understanding the origin of many different kinds of specialized terpenes remains a key goal.

Protium, a highly diverse genus within Burseraceae, has been an important model system to study the evolution of terpene metabolites. Some of the few available studies of terpene diversification have focused on *Protium* (Zapata *et al.* 2013) and have found that the coevolutionary relationship with natural enemies has an important role in shaping the chemical diversity of *Protium* species (Salazar *et al.* 2018). It has also been hypothesized that the intrinsic genetic variability that encodes the production of terpene metabolites could also facilitate the diversity of terpenoid structures in plant lineages (Pichersky *et al.* 2018). Despite the extensive knowledge of *Protium* chemistry (Rüdiger *et al.* 2018), the diversity of genes responsible for synthesizing the array of terpenoids found in the genus is unknown and few genomic resources are available.

Here, we report the transcriptome assembly and the annotation of terpene biosynthetic genes in *Protium copal*, a widespread subcanopy tree spanning environments from moist evergreen forests to seasonally deciduous forest in Central America. This novel genomic resource is now available, and we provide a table with several SSR markers that are useful for capture probe design using next-generation sequencing and identify terpene gene families that are functionally expressed in the *Protium copal* transcriptome.

MATERIALS AND METHODS

Leaf material and RNA isolation. – Mature leaves were harvested from a cultivated specimen of *Protium copal* (NY barcode and GenBank accession number will be provide during review) and stored in liquid nitrogen prior to processing. The specimen accession is available at the New York Botanical Garden. RNA isolation was performed as follows. The leaf sample was mixed with liquid nitrogen and ground into a fine powder using mortar and pestle. About 100 mg tissue powder was used for RNA extraction. Total RNAs were extracted using the Qiagen RNeasy plant mini kit. cDNA libraries were constructed for paired-end 2x100 bp HiSeq 3000 platform (Illumina, San Diego CA, USA). Total RNA volume was assessed with Qubit® 2.0 Fluorometer and purity was assessed using NanoDrop-ND 2000C spectrophotometer and bioanalyzer. The sample selected for sequencing had RNA integrity number (RINe) greater than 8.0 along with the nanodrop ratios of 1.9-2.1 (260/230) and the ratios of 2.0-2.5 (260/280). Library preparation and sequencing were performed at the QB3 Vincent J. Coates Genomics Sequencing Laboratory, University of California, Berkeley.

Transcriptome de novo assembly. – Raw reads were filtered using quality value (Q) ≥30 and demultiplexed using an option of one mismatch in index. The quality of the sequenced paired-end reads produced in each sample was observed using FastQC 0.10.1 and paired-end reads were trimmed to remove Illumina sequencing adaptors, as well as portions of poor-quality reads using the default settings Trimmomatic (Bolger *et al.* 2018). We used Trinity (Grabherr *et al.* 2011) version 2.5.1 with default parameters and a minimum contig length of 200 bp for assembly generation. Approximately, 182 million paired-end reads were used to generate the de novo assembly. Transcript abundance was analyzed using Kallisto version 0.43.1 (Bray *et al.* 2011), an alignment-free abundance estimation tool that uses pseudo-alignment to calculate transcript per million (TPM) value for each transcript. Transcript abundance was used to analyze the transcriptome assembly quality (exN50) as well as to determine the abundance of terpenoid pathway transcripts. Transcriptome

completeness was also assessed using Benchmarking Universal Single-Copy Orthologs (BUSCO) (Simão *et al.* 2011), searching for 2,121 single-copy orthologs which are conserved among eudicots. The exN50 and completeness of the transcriptome was compared with transcriptomes from two other species in Burseraceae (*Boswellia* sacra and *Bursera simaruba*).

Functional annotation. – To identify putative functions of *Protium copal* transcripts, we adopted the Trinotate pipeline to detect homology between the predicted proteins and sequences deposited in universal databases. The Trinotate pipeline uses TransDecoder v2.0.1 (Hass *et al.* 2013) to identify open reading frames (ORFs) and predict potential coding transcripts. The retrieved nucleotide sequences and putative protein sequences were then functionally annotated searching for nucleotide (BLASTx) and protein (BLASTp) homology (e-value <1e−10) against the UniProtKB/Swiss-Prot databases. Functional domains were searched against the Pfam domain database (https://pfam.xfam.org/) using HMMER v3.1b2 (Finn *et al.* 2011). The maximum e-value for reporting the best hit and associated annotation was 1e−5.

Gene ontology and COG classification. – Blast2GO v3.1 (Conesa *et al.* 2005) was additionally used to detect GO terms associated with biological processes (BPs), molecular functions (MFs), and cellular components (CCs) (Conesa *et al.* 2005, Götz *et al.* 2008). Transcripts were searched against the non-redundant (nr) database using BLASTx, which uses the translated protein sequence of each transcript as the search query (Altschul *et al.* 1990). The sequence description from the top homologous BLAST hit (e-value<10-5) was transferred to each transcript. The Clusters of Orthologous Groups (COG) consists of unique orthologous proteins that shares the same function (Tatusov *et al.* 2003). COG screening was performed using the EggNog database (Powell *et al.* 2011), integrated within the Trinotate pipeline. EggNog is a database where orthologous gene products are classified into functional categories. It is based on the principle that conserved genes are classified giving their homologous association. All transcripts where aligned to the EggNog database to predict and classify their functions.

Terpene gene diversity and phylogenetic analysis. **–** The Trinnotate and Blast2GO annotation were used to annotate genes involved in terpene, terpenoid, and isoprenoid biosynthesis, compounds which are found to make up the majority of the essential oil of species in the genus *Protium*. The GO terms that contain functional annotation related to terpene synthase activity and terpenoid biosynthetic process were identified and we use seqtk toolkit (https://github.com/lh3/seqtk) to extract the sequences matching the annotated terpene genes. To explore the evolutionary relationship between *Protium copal* terpene gene families, putative genes with identified GO terms functionally characterized up to date were used to construct a phylogenetic tree. Multiple sequence alignment of identified genes was performed using ClustalX 1.81 and the maximum likelihood tree was constructed using the MEGA 10.0.5 program and bootstrap analysis with 1,000 replicates.

Development of polymorphic SSR markers. – *Protium copal* leaf transcripts were scanned for single sequence repeat (SSR) markers using the MISA version 1.0 (Thiel *et al.* 2003). The minimum number of repeat units was defined as 10 units for mono-nucleotide repeats, 6 units for di-nucleotide, and 5 units for tri-, tetra-, penta-, and hexa-nucleotide repeats. The maximum distance between two separate repeat regions was defined to 100 bp. SNPs were discovered in the assembled transcriptome using Kissplice version 2.4.0-p1 (Sacamoto *et al.* 2012), which analyzes RNA-seq reads to identify SNPs in the sequencing library. In order to get positional data for the discovered SNPs, Transdecoder version 5.0.2 (Hass *et al.* 2013) was used to identify the open ORFs of each transcript, and BLAT

version 36x2 (Kent 2002) was used to align the identified SNPs to the transcriptome assembly, all using the default parameters.

Five random loci were selected, and primers were developed using Primer-BLAST, a tool that combines Primer3 and BLAST functionality to develop primers (Ye *et al.* 2012). Primers for two other loci found in terpene synthase genes were also developed. All primers were designed by Integrated DNA Technologies (IDT®) and the amplification was tested on a different species in the genus *Protium* (*P. heptaphyllum*, Damasco, G. 1571). We used the Thermo Scientific Phire Plant Direct PCR Master Mix designed to extract and amplify the seven SSR regions directly from leaf samples of *P. heptaphyllum*. This kit is based on Phire Hot Start II DNA Polymerase and Polymerase chain reactions (PCRs) were conducted in a total volume of 25 µl. The PCR mixture was subjected to 98 °C for 5 min, followed by 40 cycles of denaturing for 5 s at 98 °C, annealing gradient for 5 sec at 55–63 °C, extension of 20 s at 72 °C, and the final extension step at 72 °C for 1 min. PCR products of the seven SSR loci were ran on a 3% MetaPhor gel with ethidium bromide staining and purified with ExoSAP-IT® treatment (Amersham Biosciences, Buckinghamshire, UK.

RESULTS

Leaf transcriptome sequencing and de novo assembly. – llumina sequencing generated 140 GB of data containing ca.182 million paired-end reads. The Trinity assembly resulted in 78,807 transcripts with fragment sizes ranging from 224 to 12,395 bp (Figure 1a) and a median contig length of 595 bp. Table 1 presents assembly statistics of the transcriptome, including transcripts that were retained with an exN50 length of 1145 bp, which represents the shortest transcript length at which half the assembled base pairs can be found. The assembled transcriptome is considered of high quality as evidenced by the exN50 statistic and when compared to other plant transcriptome assemblies in Burseraceae (*Boswellia* sacra and *Bursera simaruba*, Figure 1b). BUSCO analysis revealed the completeness of the transcriptome assembly by searching for conserved single-copy orthologs. Of the 2,121 single-copy orthologs, 1,2,92 (60.9%) were found complete, of which 563 (26.5%) were duplicated, 515 (24.3%) were found but fragmented, and 314 (14.8%) were missing. Of the three Burseraceae transcriptomes, *Protium copal* was the most complete, with the fewest missing and fragmented orthologs (Figure 1c).

Functional annotation and similarity with other plant genomes. – According to Trinotate annotation, 38,042 nucleotide sequences (48.27%) and 24,250 protein sequences (30.77%) displayed significant homology with Viridiplantae when aligned against the UniProtKB/Swiss-Prot database using BLASTx and BLASTp searches, respectively. Furthermore, 20,271 (25.72%) unique Pfam protein motifs were assigned. These protein domains can be involved in various biological processes and molecular function such as protein–protein interactions, transcription regulation, and organic compound biosynthetic processes. In contrast, Blast2GO annotation retrieved 48,951 transcripts (69.91%) were successfully matched to homologous sequences for Viridiplantae in the NR database (e-value < 1e-5). Out of the blasted transcripts, 18,822 transcripts (26.87%) were mapped to at least one gene ontology (GO) term using Blast2GO. The vast majority of BLAST hits for the transcripts came from the genus *Citrus*, also a member of the order Sapindales (Figure 2a). Our transcript dataset displayed 26,223 (37.44%) hits with the NR database and to *Citrus* sinensis, followed by *Citrus* clementina (10,185 top-hits, 14.54%), Theobroma cacao (2,010 top-hits), and Hevea brasiliense (1,623 top-hits) and Vitis vinifera (1,484 top-hits) (Figure 2A).

Gene ontology (GO) and cluster of orthologous groups (COG). **–** GO analysis revealed 4,532 unique GO terms related to plant gene ontology. Among the three main categories, Biological processes (BP) category was the most abundant (2,340 GOs), followed by molecular function (MF, 1,750 GOs) and cellular component (CC, 104 GOs) categories (Figure 2b). Within the BP category, metabolic processes (34.02%), cellular process (29.07%), and single-organism processes (16.28%) were most represented. Likewise, genes encoding binding proteins (33.40%) and genes encoding proteins related to catalytic activities (31.95%) were most abundant in the MF category. In the CC category, membrane (22.26%), membrane part (18.78%), cell (16.70%), and cell part (16.49%) were abundantly represented GO terms. In total, 10,960 (45% of the transcripts with NR blast hits for Viridiplantae) transcripts were assigned to different COG functional categories. The largest group is represented by the serine theorine protein kinase (COG0515, 2,014 hits, 18%), followed by leucine rich repeat (COG4886, 491 hits, 13%), ankyrin repeat (COG0666, 192 hits). A few other clusters, such us chromatin structures and dynamics, cell motility, extracellular structures are underrepresented or absent.

Diversity and phylogeny of terpene biosynthetic genes. – A total of 68 transcripts were identified as being part of monoterpenoid biosynthetic processes (GO:0016099), triterpenoid biosynthetic processes (GO:0016104), including pentacyclic (GO:0019745) and tetracyclic triterpenoids (GO:0010686), terpenoid transport (GO:0046865), terpene synthase activity (GO:0010333) and terpene/terpenoid biosynthetic processes (GO:0046246 and GO:0016114, respectivelty). The lengths of these transcripts varied between 232 bp and 4278 bp, with a median contig length of 736 bp. In total, we found approximately 540.2 transcripts per million (TPM) expressed for all terpene and terpenoid functional annotation (Figure 3). Genes responsible for the synthesis of monoterpenoids were relatively highly expressed with a single annotated gene having 58.7 TPM. In addition, the phylogenetic analysis showed that monoterpenoid putative genes are closely related to the triterpenoid gene subfamily in *Protium copal*. Regarding different molecule structure, tetracyclic and pentacyclic triterpenoids are also phylogenetically more similar than other putative genes. Although, additional terpene gene subfamilies were not identified in the annotation analysis, the phylogeny suggests strong genetic structure within general terpene and terpenoid biosynthetic putative genes. There are at least two distinctive unidentified terpene clades with differential phylogenetic structure that are potentially expressed in *Protium copal* transcriptome.

Single sequence repeat (SSR) identification. – A total of 11,480 repeat regions were identified in the transcriptome, found in 9,496 transcripts (15%). In addition, 1610 transcripts contained multiple SSR's (2.5%), and a total of 770 compound SSR's were identified in the *P. copal* transcriptome. Out of the 11, 480 total SSR regions, there were 7,308 mono-nucleotide (64%), 1,936 di-nucleotide (17%), 2,010 tri-nucleotide (18%), 112 tetra-nucleotide (1%), 53 penta-nucleotide (0.5%), and 61 hexa-nucleotide regions (0.5%). Within transcripts involved in terpenoid biosynthetic pathways, 10 SSR regions were identified, of which there are 5 mono-nucleotide, 2 di-nucleotide, 1 tri-nucleotide, and 2 compound SSR regions. Out of the 10 transcripts containing SSR regions, 7 are involved in the isoprenoid biosynthetic pathway and 3 are involved in the terpenoid biosynthetic pathway.

SNP discovery and primer validation. – A total of 64,510 SNPs was identified among 25,505 transcripts, of which 36,893 are located in coding regions (CDS). 22,292 SNPs were classified as non-synonymous (35%), of which 22,159 are located in CDS regions. Of the 64,510 SNPs, 74

SNPs were located in transcripts involved in terpenoid biosynthesis pathway, located in 30 transcripts. Of these SNPs, 28 were located in CDS regions (38%). 14 SNPs were classified as non-synonymous, all of which were located in CDS regions. Of the 74 identified SNPs, 57 were involved in isoprenoid biosynthesis genes, 16 in terpene synthase genes, and 1 in terpenoid biosynthesis genes. Primers were developed for five randomly polymorphic loci, as well as two polymorphic loci located in terpene synthase genes in order to validate the in-silico SNP detection, and create primers for further study of terpene synthase genes (Table 2). PCR amplification of *Protium* heptaphyllum, a congeneric species, was successful performed for all primers and the agarose gel electrophoresis of the PCR product yielded a strong band of correct target fragment size for the seven developed primers (Figure S1).

DISCUSSION

Novel transcriptomic resources for Burseraceae. – Although Burseraceae is distributed throughout tropical and subtropical regions of the world, the majority of the transcriptomic information available is limited to Asiatic and African genera, such as *Boswellia* (frankincense), and *Bursera* (linaloe) occurring in Central America (Matasci *et al.* 2014). Here, we generate the first transcriptome annotation of *Protium*, a remarkable genus that harbors over 170 plant species in the Neotropics (Daly *et al.* 2012, Fine *et al.* 2014). *Protium* is globally known for the aromatic and medicinal properties of resins and essential oils (Rüdiger *et al.* 2018). The availability of a comprehensive leaf transcriptome for *Protium copal* is the first step toward the development of genomics and medical applications. The transcriptomic data generated in this study provide useful resources to explore the functional aspects of Burseraceae resinous chemicals and investigate their genetic associations. Enzymes responsible for the synthesis of various secondary metabolites described in this species may be identified from the provided set of annotated gene domains, assembled transcripts or even based on the raw sequencing reads.

In addition, *Protium copal* transcripts could be used to generate novel sequencing markers applied to population genetics and comparative phylogenetics studies. High-throughput sequencing based on transcriptome capture are designed to enrich target genomic regions (Bi *et al.* 2012) and these techniques have been commonly used to generate well-resolved phylogenies for tropical plant groups (Prata *et al.* 2018, Sass *et al.* 2016]. *Protium* is considered a monophyletic genus and one of the most dominant plant genera in the Amazon region (Fine *et al.* 2014). The development of target sequencing regions based on the annotated transcriptome of *Protium copal* could be directly used to improve the estimates of species limits and genetic diversity within *Protium*. Further transcriptomic analysis on multiple tissues and specimens, rather than a single reference, is still necessary to circumvent sampling bias and ensures that the majority of genetic diversity within *Protium copal* is fully captured. Furthermore, de novo construction of a pan-transcriptome for *Protium* could help to unravel variation in gene regulation and expression and provide additional candidate genes for the study of select genotypes.

Assembly and annotation quality. – We found that the assembly quality of *P. copal* transcriptome is equivalent to other transcriptomes of plant species in Burseraceae (Matasci *et al.* 2014). Over 50% of the transcripts and predicted proteins were successfully assigned to genes by BLASTx and BLASTp searches and more than 75% of transcripts returned a homologous BLAST hit with an e-value < 1e-5, an indication of the high assembly quality for *Protium copal* transcriptome. The exN50 is calculated similarly to N50 used for genome assemblies except that

it is limited to the top most highly expressed transcripts that represent a 50% of the total normalized expression data. In contrast to whole-genome assemblies, transcriptomes might not achieve contigs with high exN50 values and the most highly expressed transcripts may not be the longest ones. However, exN50 for *Protium copal* is superior to other transcriptome assemblies in Burseraceae and our statistics are comparable to recently published high-quality plant transcriptomes (He *et al.* 2012, Tian *et al.* 2015, Li *et al.* 2016). On the other hand, BUSCO represents a more appropriate measure to access transcriptome quality by quantifying the presence of conserved orthologs in an assembly (Simão *et al.* 2011). In comparison to transcriptomes in Burseraceae (*Boswellia sacra* and *Bursera simaruba*), the *Protium copal* transcriptome had the least amount of missing single-copy orthologs, indicating a relatively complete assembly, and ~60% of orthologs were found in the assembly, indicating a relative high-quality assembly.

Detection and validation of SSR markers in *Protium copal*. – SSR markers are frequently designed from transcriptomic assemblies providing a suitable source for genetic diversity assessment via low-cost projects and straightforward bioinformatic pipelines. Despite being derived from coding DNA regions, which tend to be evolutionarily conserved, SSRs developed from transcriptomes are considered a valued genomic resource for studying the genetic structure of plant populations (Taheri *et al.* 2018). Here, we have made available a database of SSR markers known to be polymorphic within *Protium copal* that will be useful for genetic studies on this important medicinal plant. This database presents a large collection of putative expressed genes that can be used in further genetic linkage and genome association analysis. In addition, primers can be easily designed for targeting specific genes in *Protium* and closely related genomes. In our PCR validation, all 42 accessions of *Protium* heptaphyllum, a congeneric plant species from South America, were successfully amplified. The sequencing validation of SSRs designed in this study is the next step to provide a marker set for genetic development efforts.

High similarity with *Citrus* genome. – The number of putative genes identified in the *Protium copal* transcriptome was within the range of putative functional genes identified in other angiosperms (Falara *et al.* 2011, Nieuwenhuizen *et al.* 2013, Külheim *et al.* 2015). As expected, the percent identity of assembled transcriptomes tends to increase as comparisons are made with plants at the same order, family or genus level (Geniza *et al.* 2017). Within our transcriptome assembly, over 38% of the transcripts were successfully annotated against one or more species of *Citrus*, a plant crop with important genomic resources. *Citrus* is a domesticated plant within the Rutaceae family (Wu *et al.* 2018), and likewise Burseraceae, is a member of the Sapindales order. The *Citrus* genome database is an open-source genome funded by the USDA and NSF agencies to enable basic and applied genomics, genetics, breeding and disease research (Xu *et al.* 2013). Plant families in the Sapindales (e.g. Anacardiaceae, Burseraceae, Rutaceae) are known for producing a diverse suite of aromatic chemicals. Terpene gene families are found in species with specialized structures for storing volatile terpenes (Külheim *et al.* 2015), such as *Citrus*, which accumulate volatiles in oil glands. Our transcriptome annotation includes an increased complement of coding genes also found in *Citrus* that are useful to identify proteins involved in the different steps of the terpene biosynthesis pathway.

Diversity of terpene biosynthetic genes. – Thousands of different terpenoid compounds are produced by plants through the expression of terpene synthase (TPS) genes. Terpenoids are characterized by an isoprenoid chemical structure and include derivatives with various functional groups (Pichersky *et al.* 2018). The TPS gene family is classified according to phylogenetic

relationships into eight subfamilies which comprise mono-, sesqui-, di- and triterpene synthases (Nagegowda 2010). In this study, the annotation the *P. copal* transcriptome revealed high diversity of terpenoid genes with different biosynthetic pathways (e.g. monoterpenoids and triterpenoid) and genes responsible for terpene cell transportation. Terpenoid synthase activity expressed in *P. copal* is primarily responsible for the synthesis of linear terpenes (e.g. isopentenyl-PP, geranyl-PP, farnesyl-PP and geranylgeranyl-PP) containing varying numbers of isoprene units. Triterpenoids genes annotated in *P. copal* regulates the chemical reactions and pathways that result in six isoprene units and 4 or 5 carbon rings (tetracyclic and pentacyclic synthases). In addition, TPS genes found in *P. copal* results in the formation of monoterpenoids having a C10 molecule skeleton. We also found genes that are directly associated with the movement of terpenoids into or between conduit cells.

The origin and evolution of plant secondary metabolites in Burseraceae are an emerging theme in plant phytochemistry (Salazar *et al.* 2018). Although there are a few hundred terpenoids produced by almost all plants, the vast majority of terpenoids are restricted to a given lineage, or even a single species, and new terpenoids keep arising in many plants (Pichersky *et al.* 2018). Recent studies have not yet provided answers to how terpenoids became the largest class of compounds and often the largest class of specialized compounds produced by plants. The coevolutionary arms race hypothesis (Ehrlich and Raven 1964) has been thought for a long time to explain the diversity of terpene metabolites in aromatic plants like Burseraceae, in which specialized metabolites are predicted to diversify as a response of counter-defense between plants and specific herbivores. Defense against biological enemies is a well-recognized function of plant terpenoids and the production of terpenes represents a major evolutionary advantage against natural enemies (Kessler *et al.* 2011).

In addition, the presence of a large number of genes already known to be involved in terpene biosynthesis could also underlie the ability to synthesize large numbers of terpenoids and increases the probability that new terpene genes will evolve. As a result of the complex chemical and physical properties of terpene metabolites, the diversity of terpenoids found across linages may also be driven by other factors besides herbivore defense due to their role in primary metabolic functioning (e.g. electron transport chains depend on terpene association) and reproduction (e.g. chemical signaling to pollinators) (Pichersky *et al.* 2018). Although examples of diverse terpenoid compounds have been found in some model organisms, the study of their evolution requires comparison with closely related species that occupy different ecological niches. Here, we provided a useful transcriptome annotation of terpene genes for *Protium* that will be relevant to study the evolution and diversification of terpene secondary metabolites including multiple related lineages in Burseraceae, one of the largest resinous plant families that produce a diverse array of terpene-related chemistry.

REFERENCES

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

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., Good, J. M. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC genomics, 2012, 13(1), 403.

Bolger, A. M., Marc L., Bjoern U. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 2014, 30(15), 2114-2120.

Bray, N. L., Pimentel, H., Melsted, P., Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nature biotechnology, 34(5), 525.

Chen, F., Tholl, D., Bohlmann, J., Pichersky, E. The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. The Plant Journal, 2011, 66(1), 212-229.

Cheng, A. X., Lou, Y. G., Mao, Y. B., Lu, S., Wang, L. J., Chen, X. Y. Plant terpenoids: biosynthesis and ecological functions. Journal of Integrative Plant Biology, 2007, 49: 179-186.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., Robles, M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics, 2005, 21(18), 3674-3676.

Daly, D. C., Fine, P. V. A., Martínez-Habibe, M. C. Burseraceae: a model for studying the Amazon flora. Rodriguésia, 2012, 63(1), 021-030.

Ehrlich, P. R., Raven, P. H. Butterflies and plants: a study in coevolution. Evolution, 1964, 18(4), 586-608.

Falara V., Akhtar T. A., Nguyen T., Spyropolou E. A., Bleeker P. M., Schauvinhold I. *et al.* The tomato terpene synthase gene family. Plant Physiol., 2011, 157, 770–789.

Fine, P. V. A., Zapata, F., Daly, D. C. Investigating processes of neotropical rain forest tree diversification by examining the evolution and historical biogeography of the Protieae (Burseraceae). Evolution, 2014, 68(7), 1988-2004.

Finn, R. D., Clements, J., Eddy, S. R. HMMER web server: interactive sequence similarity searching. Nucleic acids research, 2011, 39, 29-37.

Geniza, M., Jaiswal, P. Tools for building de novo transcriptome assembly. Current Plant Biology, 2017, 11, 41-45.

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic acids research, 2008, 36(10), 3420-3435.

Grabherr, M.G., Haas, B.J., Yassour, M. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011, 29(7), 644-52.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols, 2013, 8(8), 1494.

He, R., Kim, M. J., Nelson, W., Balbuena, T. S., Kim, R., Kramer, R. *et al.* Next-generation sequencing-based transcriptomic and proteomic analysis of the common reed, Phragmites australis (Poaceae), reveals genes involved in invasiveness and rhizome specificity. American Journal of Botany, 2012, 99(2), 232-247.

Kent, W. J. BLAT—the BLAST-like alignment tool. Genome research, 2002, 12(4), 656-664.

Kessler, A., Heil, M. The multiple faces of indirect defences and their agents of natural selection. Functional Ecology, 2011, 25: 348–357.

Külheim C., Padovan A., Hefer C., Krause S. T., Köllner T. G., Myburg A. A. *et al.* The Eucalyptus terpene synthase gene family. BMC Genomics, 2015, 16, 450–466.

Langenheim, J. H. Plant resins: chemistry, evolution, ecology, and ethnobotany. No. 620.1924 L275p. Oregon, US: Timber Press, 2003.

Li, M., Liang, Z., Zeng, Y., Jing, Y., Wu, K., Liang, J. *et al.* De novo analysis of transcriptome reveals genes associated with leaf abscission in sugarcane (Saccharum officinarum L.). BMC genomics, 2016, 17(1), 195.

Matasci, N., Hung, L. H., Yan, Z., Carpenter, E. J., Wickett, N. J., Mirarab, S., *et al.* Data access for the 1,000 Plants (1KP) project. Gigascience, 2014, 3(1), 17.

Nagegowda, D. A. Plant volatile terpenoid metabolism: biosynthetic genes, transcriptional regulation and subcellular compartmentation. FEBS letters, 2010, 584(14), 2965-2973.

Nieuwenhuizen N. J., Green S. A., Chen X., Bailleul E. J. D., Matich A. J., Wang M. Y., *et al.* Functional genomics reveals that a compact terpene synthase gene family can account for terpene volatile production in apple. Plant Physiol., 2013, 161, 787–804.

Pichersky, E., Raguso, R. A. Why do plants produce so many terpenoid compounds? New Phytologist, 2018, 220(3), 692-702.

Powell, S., Szklarczyk, D., Trachana, K., Roth, A., Kuhn, M., Muller, J., *et al.* EggNOG v3. 0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic acids research, 2011, 40, 284-289.

Prata, E. M., Sass, C., Rodrigues, D. P., Domingos, F. M., Specht, C. D., Damasco, G., *et al.* Towards integrative taxonomy in Neotropical botany: disentangling the Pagamea guianensis species complex (Rubiaceae). Botanical Journal of the Linnean Society, 2018, 188(2), 213-231.

Rüdiger, A. L., Siani, A. C., Junior, V. V. The chemistry and pharmacology of the South America genus *Protium* Burm. f. (Burseraceae). Pharmacognosy reviews, 2007, 1(1), 93-104.

Sacomoto, G. A., Kielbassa, J., Chikhi, R., Uricaru, R., Antoniou, P., Sagot, M. F., ... & Lacroix, V. (2012, December). K is splice: de-novo calling alternative splicing events from rna-seq data. In BMC bioinformatics (Vol. 13, No. 6, p. S5). BioMed Central.

Salazar, D., Lokvam, J., Mesones, I., Vásquez, M., Ayarza, J. M., Fine, P. V. A. Origin and maintenance of chemical diversity in a species-rich tropical tree lineage. Nature ecology & evolution, 2018, 2(6), 983-990.

Sass, C., Iles, W. J., Barrett, C. F., Smith, S. Y., Specht, C. D. Revisiting the Zingiberales: using multiplexed exon capture to resolve ancient and recent phylogenetic splits in a charismatic plant lineage. PeerJ, 2016, 4, e1584.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 2015, 31(19), 3210-3212.

Stacey, R. J., Cartwright, C. R., McEwan, C. Chemical characterization of ancient mesoamerican copal resins: preliminary results. Archaeometry, 2006, 48(2), 323-340.

Taheri, S., Lee Abdullah, T., Yusop, M., Hanafi, M., Sahebi, M., Azizi, P., Shamshiri, R. Mining and development of novel SSR markers using next generation sequencing (NGS) data in plants. Molecules, 2018, 23(2), 399.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., *et al.* The COG database: an updated version includes eukaryotes. BMC bioinformatics, 2003, 4(1), 41.

Thiel, T., Michalek, W., Varshney, R., Graner, A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). Theoretical and applied genetics, 2003, 106(3), 411-422.

Tian, X. J., Long, Y., Wang, J., Zhang J. W., Wang, Y. Y., Li, W. M., *et al.*. De novo transcriptome assembly of common wild rice (Oryza rufipogon Griff.) and discovery of drought-response genes in root tissue based on transcriptomic data. PLoS One, 2015, 10(7), e0131455.

Wu, G. A., Terol, J., Ibanez, V., López-García, A., Pérez-Román, E., Borredá, C. *et al.* Genomics of the origin and evolution of Citrus. Nature, 2018, 554(7692), 311.

Xu, Q., Chen, L. L., Ruan, X., Chen, D., Zhu, A., Chen, C. *et al.* The draft genome of sweet orange (Citrus sinensis). Nature genetics, 2013, 45(1), 59.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., Madden, T. L. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC bioinformatics, 2012, 13(1), 134.

Zapata, F., Fine, P.V.A. Diversification of the monoterpene synthase gene family (TPSb) in *Protium*, a highly diverse genus of tropical trees. Molecular Phylogenetics & Evolution, 2013, 68: 432–442.

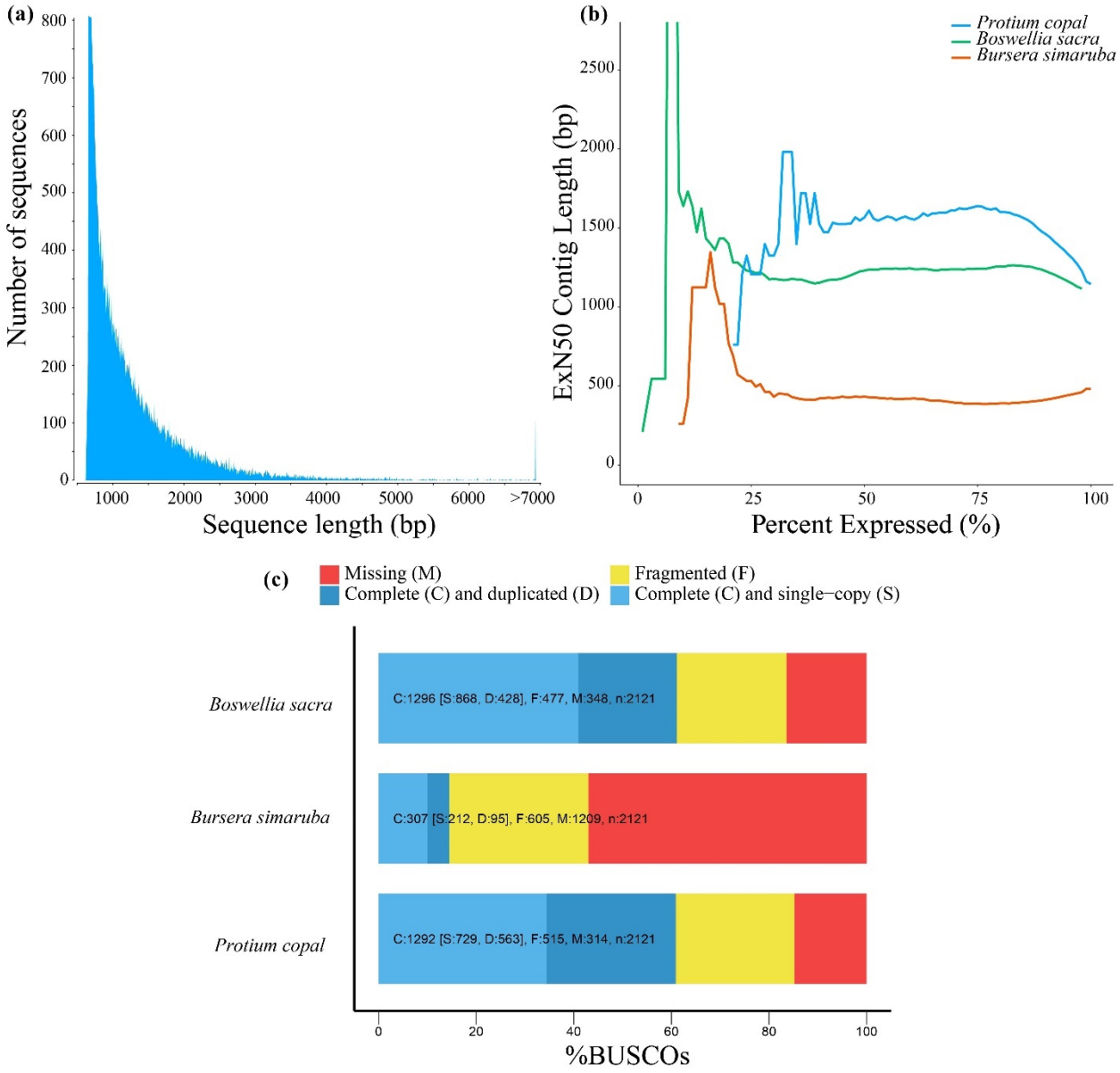

Figure 1. (A) Diagram of sequence length for the transcriptome assembly of Protium copal (Burseraceae). (B) The N50 value, the shortest transcript length at which 50% of assembled bases can be found, calculated only for the top percentile of expressed transcripts. The maximum exN50 value is at 75% expression, with an N50 of 1639. (C) Benchmarking Universal Single-Copy Orthologs (BUSCO) results of the Protium copal transcriptome assembly in comparison to two other transcriptomes in Burseraceae (Bursera simaruba and Boswellia sacra).

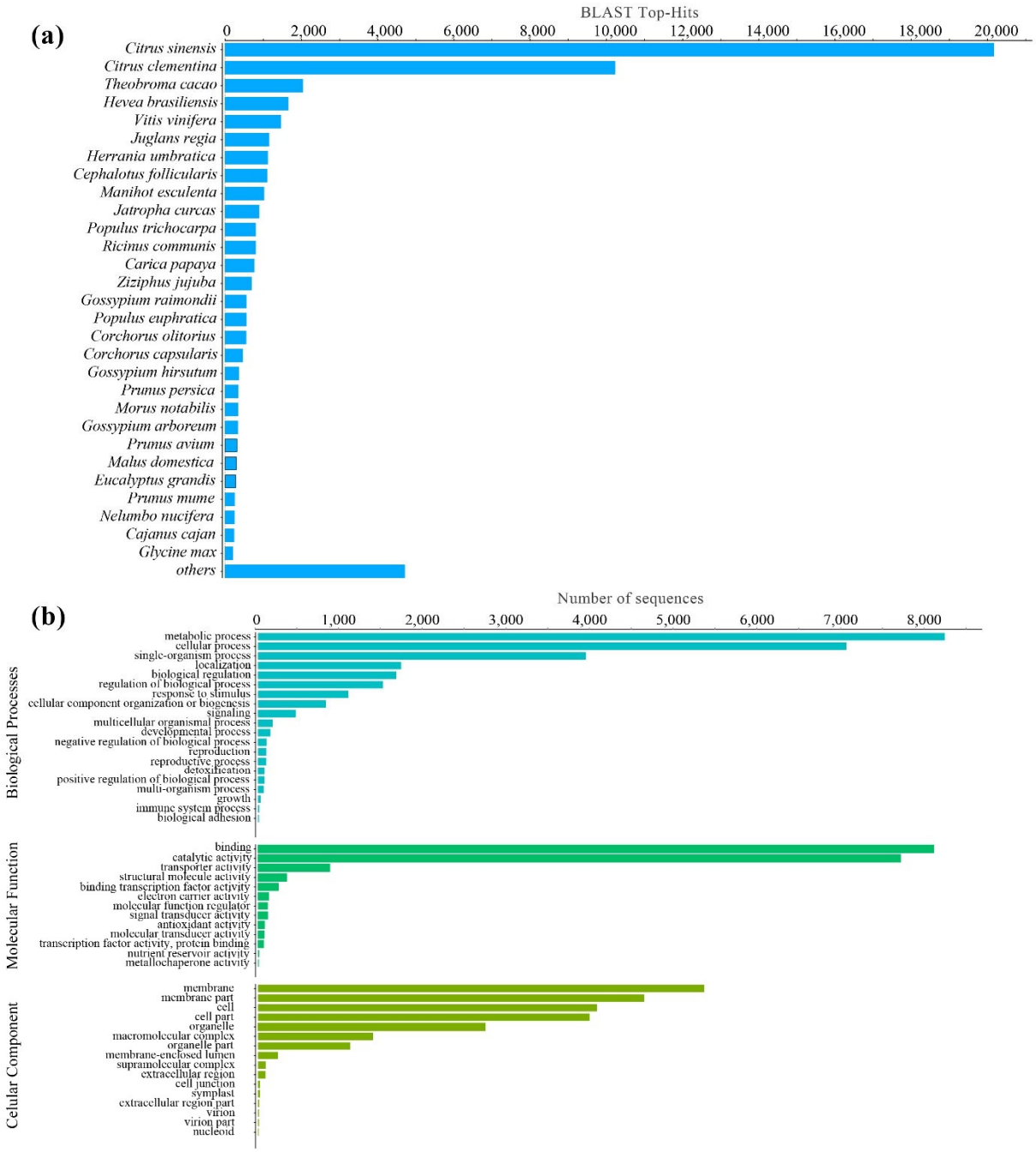

Figure 2. (A) The species distribution of BLAST hits used to assign sequence descriptions and Gene Ontology terms to transcripts for annotation. The majority of transcripts had BLAST hits from the genus Citrus, a member of the same order Sapindales. (B) Top 20 Gene Ontology terms (Level 2) for each of the three main sub-categories, Molecular Function (MF), Biological Process (BP), and Cellular Component (CC), based on the number of transcripts assigned that GO term.

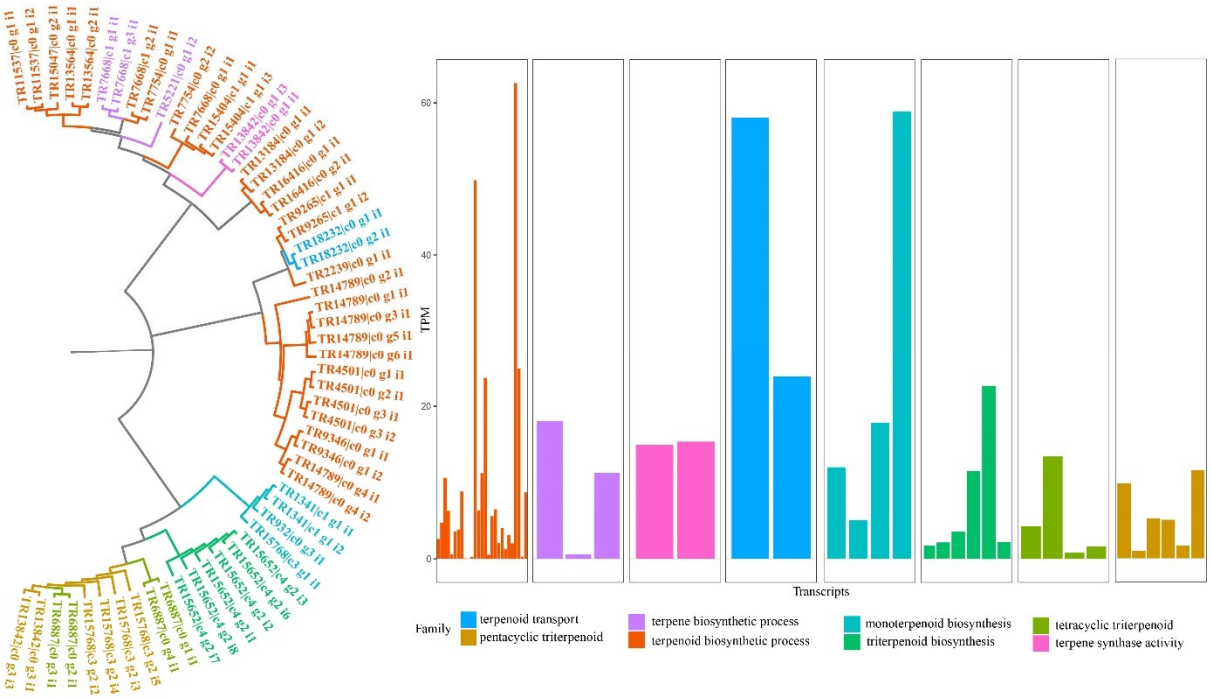

Figure 3. Maximum likelihood tree showing the phylogenetic relationship among putative terpene genes annotated in the Protium copal (Burseraceae) transcriptome and the expected abundance of transcripts expressed in transcripts per million (TPM). TPM is the abundance one would expect to find in a pool of a million transcripts, so essentially the per million expressed transcripts.

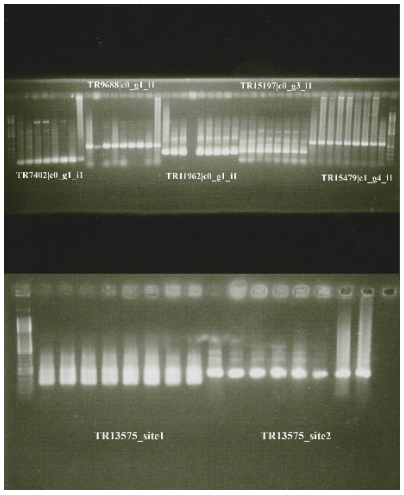

Figure S1. Agarose gel electrophoresis of the PCR products from the congeneric species *Protium heptaphyllum* successful amplified for seven SSR markers designed using the *Protium copal* transcriptomic resources. The PCR mixture was subjected to 98 °C for 5 min, followed by 40 cycles of denaturing for 5 s at 98 °C, annealing gradient for 5 sec at 55–63 °C, extension of 20 s at 72 °C, and the final extension step at 72 °C for 1 min.

Table 1. Trinity transcriptome assembly statistics for *Protium copal* (Burseraceae).

Total trinity genes	**44,754**
Total trinity transcripts	63,288
Percent GC	41.56
Contig N10	2,897
Contig N20	2,102
Contig N30	1,696
Contig N40	1,396
Contig N50	1,145
Median contig length:	595
Average contig	832.27
Total assembled bases	52,672,403

Table 2. Primer sequences designed for *Protium copal* based on the developed SSR markers that were used for PCR validation.

Primer ID	Forward Primer	Reverse Primer	Blast Description
TR13575_site1	TCTGTCCCGTGCA AAACTCA	TTGGAGGCAAGG TGGTTCAT	terpene synthase 10-like
TR13575_site2	GCTTGACGTACTC CTTGAGGT	TGACTTGCAACG ACTTGGAG	terpene synthase 10-like
TR7402|c0_g1_i1	CTTTCATGGACGC ACCAACG	TCAAGGTTCGCA ACTGGGTT	hypothetical protein CICLE_v10016425mg
TR9688|c0_g1_i1	AGGGCGTCAATTG AGTACTGG	TTATCCATGTTTG GGCCTGGG	hypothetical protein CISIN_1g027549mg
TR11962|c0_g1_i1	GTCCTTTGGCTCTG CGTTTG	AGAAGTGCGTGA TGCTCGAA	DAR GTPase chloroplastic
TR15197|c0_g3_i1	AGCCACAAGACA ACAGCTCC	GAAGGGCTCATC GCATCTGA	hypothetical protein POPTR_0003s10040g
TR15479|c1_g4_i1	GGTGGCATGCCCT TAGTCAA	GAGGCTACTTCA CATGGCGT	Gag-protease-integrase-RT-R poly