RSC Advances

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2021, **11**, 25096

UHPLC-HRMS/MS on untargeted metabolomics: a case study with *Copaifera* (Fabaceae)†

Ananda da Silva Antonio,[ab] Davi Santos Oliveira,[a] Gustavo Ramalho Cardoso dos Santos,[b] Henrique Marcelo Gualberto Pereira,[b] Larissa Silveira Moreira Wiedemann[a] and Valdir Florêncio da Veiga-Junior*[ac]

Untargeted metabolomics is a powerful tool in chemical fingerprinting. It can be applied in phytochemistry to aid species identification, systematic studies and quality control of bioproducts. This approach aims to produce as much chemical information as possible, without focusing on any specific chemical class, thus, requiring extensive chemometric effort. This study aimed to evaluate the feasibly of an untargeted metabolomics method in phytochemistry by a study case of the *Copaifera* genus (Fabaceae). This genus contains significant medicinal species used worldwidely. *Copaifera* exploitation issues include a lack of chemical data, ambiguous species identification methods and absence of quality control for its bioproducts. Different organs of five *Copaifera* species were analysed by UHPLC-HRMS/MS, GNPS platform and chemometric tools. Untargeted metabolomics enabled the identification of 19 chemical markers and 29 metabolites, distinguishing each sample by species, plant organs, and biome type. Chemical markers were classified as flavonoids, terpenoids and condensed tannins. The applied method provided reliable information about species chemodiversity using fast workflow with little sampling size. The untargeted approach by UHPLC-HRMS/MS proved to be a promising tool for species identification, pharmacological prospecting and in the future for the quality control of extracts used in the manufacture of bioproducts.

Received 23rd April 2021
Accepted 17th June 2021

DOI: 10.1039/d1ra03163e

rsc.li/rsc-advances

Introduction

Plants have always been present in humankind history as a reliable and vast source of medicine and drugs against several diseases, such as cancer, inflammations, and viruses.[1] The exploitation of medicinal herbs must be carefully performed to always obtain reproducible and reliable bioproducts. Even small changes in the chemical composition can modify the biological properties of a natural extract. Such changes in a plant's chemical profile can be due to intraspecific variations within a variety or species, abiotic stress, geographic influences in plant development, or even a wrong identification of species during logging.[2–4] A drawback in medicinal herb exploitation is that the scientific knowledge of them is still scarce compared to the magnitude of this economic and historical importance, as a homemade traditional medicine and as a recognized option in modern medicine treatments.

In order to ensure efficacy, reproducibility, and biodiversity, preservation and quality control, methods able to detect slight changes in the chemical profile must be addressed.[1] Untargeted metabolomics come forward as an alternative to enhance botanic identification and natural products quality control. This research field aims to generate a holistic view of sample chemical composition, without focusing on any specific chemical class. Thus, it is reliable in the exploitation of new matrix, elucidation of unusual chemical markers and screening for new bioactive metabolites.[5–7]

The holistic chemical profile is usually achieved by high performance exploratory methods, such as chromatographic and spectrometric techniques, as they can provide a huge set of chemical data with compounds identification without the need of time-consuming isolation methods.[5,6] In addition, an extensive chemometric analysis is required on untargeted metabolomic studies.[8,9] The analysis of comprehensive chemical profiles by chemometric methods can address models to achieve botanic species identification, distinction of already harvested trees, increase the biodiversity knowledge, quality control of bioproducts, phytochemical screening for bioactive compounds and

[a]*Chemistry Department, Institute of Exact Sciences, Federal University of Amazonas, Avenida Rodrigo Octávio, 6200, Coroado, CEP: 69.077-000, Manaus, AM, Brazil. E-mail: valdir.veiga@gmail.com*

[b]*Federal University of Rio de Janeiro, Chemistry Institute, Brazilian Doping Control Laboratory (LBCD – LADETEC/IQ – UFRJ), Avenida Horácio Macedo, 1281 - Pólo de Química - Cidade Universitária, Ilha do Fundão, CEP: 21941-598, Rio de Janeiro, RJ, Brazil*

[c]*Chemical Engineering Section, Military Institute of Engineering, Praça General Tibúrcio, 80, Praia Vermelha, Urca, CEP: 22.290-270, Rio de Janeiro, RJ, Brazil*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d1ra03163e

Table 1 Specialized metabolites detected (D) in *Copaifera* samples by high resolution mass spectrometry (HRMS)[a]

Identification number	Detected *m/z*	*m/z* error (ppm)	Retention time (min.)	Identity	*C. multijuga*			*C. venezuelana*		*C. langsdorfii*				*C. trapezifolia*	*C. lucens*	
					Branch	Leaf	Bark	Branch	Bark	Seed	Bark	Flower	Leaf	Leaf	Leaf	Bark
1	329.066	−2.05	7.10	3,7-Di-*O*-methylquercetin	D	—	—	—	—	—	—	—	D	—	D	—
2	317.212	−0.68	8.41	2-Oxo-kolavenic acid	D	D	—	D	—	—	—	—	—	—	—	—
3	307.191	−1.57	7.14	5-*O*-Methylembelin	—	—	—	—	—	—	—	—	—	D	—	—
4	333.207	−0.39	7.36	6β,7β-Dihydroxykaurenoic acid	D	D	D	—	—	—	—	—	—	—	—	D
5	313.238	−1.38	8.47	7,8-Dihydroxyoctadecenoic acid	—	—	—	—	—	D	—	—	—	—	—	D
6	295.227	−2.90	8.41	8-Hydroxyoctadeca-9,12-dienoic acid	—	—	—	—	—	D	—	D	—	—	—	—
7	281.248	−2.14	10.00	9-Octadecenoic acid	—	—	—	—	—	D	—	—	—	—	—	—
8	431.098	−0.85	5.92	Afzelin	—	—	—	—	—	—	—	—	D	—	D	—
9	299.056	−0.37	7.04	Chrysoeriol	D	—	—	—	—	—	—	—	D	—	—	—
10	313.072	0.76	7.49	Cirsimaritin	—	—	—	—	—	—	—	—	D	—	—	D
11	335.223	0.64	8.38	Ent-16β,17-dihydroxy-19-kauranoic acid	D	D	—	D	—	—	—	—	—	—	—	—
12	833.209	0.35	4.67	Ent-fisetinidol-(4β->8)-catechin-(6->4β)-*ent*-fisetinidol	—	—	D	—	D	—	—	—	—	—	—	D
13	561.140	−0.41	4.31	Epiafzelechin-(4β->8)-catechin	—	—	D	—	D	—	—	—	—	—	—	D
14	343.082	−0.95	7.01	Eupatilin	—	—	—	—	—	—	—	—	D	—	—	—
15	243.029	−3.69	5.52	Gentisein	—	—	—	—	—	—	—	—	—	—	—	D
16	285.040	−1.62	6.05	Luteolin	—	—	—	—	—	—	D	—	D	D	—	—
17	455.353	−0.15	9.47	Ursolic acid	—	—	—	—	D	—	D	—	—	—	—	D
18	271.061	0.34	5.20	3′,4′,7-Trihydroxyflavanone	—	—	—	D	—	—	—	—	—	—	—	—
19	495.078	0.43	3.91	3,4-Di-*O*-galloylquinic acid	—	—	—	—	—	—	—	—	—	—	D	—
20	329.067	1.76	6.93	4,6-Dihydroxy-2-[(4-hydroxy-3,5-dimethoxyphenyl)methylidene]-1-benzofuran-3-one	D	—	—	—	—	—	—	—	—	—	—	—
21	269.045	−2.03	6.00	Apigenin	—	—	—	D	—	—	—	—	—	—	—	—
22	624.136	4.48	4.44	Cyanidin 3-*O*-(2-*O*-β-D-glucuronosyl)-β-glucoside	—	—	D	—	D	—	—	—	—	—	—	—
23	319.046	0.18	5.04	Dihydromyricetin	—	—	—	—	—	—	—	—	—	D	—	—
24	561.140	−38.07	3.99	Flavonol 3-*O*-β-glucosyl-(1->2)-β-glucoside	—	—	D	D	D	—	—	—	—	—	—	—
25	195.050	−5.26	0.33	Gluconic acid	—	D	—	—	—	—	—	—	—	—	—	—
26	285.040	−1.62	6.17	Luteolin	D	—	—	D	—	—	D	—	—	D	—	D
27	301.035	0.30	5.92	Quercetin	—	—	—	—	—	—	—	—	—	D	—	—
28	447.093	0.27	5.32	Quercitrin	—	—	—	—	—	—	—	—	D	D	D	—
29	191.056	4.71	0.33	Quinic acid	—	D	—	—	—	—	—	—	—	—	D	—

[a] Not detected: —.

pharmacological prospection in several types of sample, ranging from raw botanic material to commercial bioproducts.[5,7,10,11]

The genus *Copaifera* (Fabaceae, Caesalpinoideae) is an example of a recognized medicinal plant which still faces difficult on its commercial and scientific exploitation. *Copaifera* species are famous medicinal plant across the world, popular known as "Copaíba". Copaiba trees produce and exude the copaiba oil, an oilresin with remarkable ethnobotanic usage and applications in the manufacturing of several bioactive products.[12,13] *Copaifera* oilresins has been studied to their extraordinary bioactivities, such as, antioxidants, anti-inflammatory, cytotoxic and gastroprotector.[14] These pharmacological properties are mainly related to its sesquiterpenes (*e.g.* α-copaene, α-humulene and caryophyllene oxide) and diterpenes (*e.g.* copalic acid and *ent*-agatic acid) chemical composition.[15] In addition to the oilresin, non-volatile extracts of copaiba twigs, branches, barks and trunk trees have several therapeutical applications in traditional medicine. However, copaíba extracts are not as famous as the oilresins, and present substantially few scientific exploitations.[16–18] Several review papers have already described copaíba oilresin composition and pharmacological properties.[19,20] Nonetheless, the composition of the extracts from *Copaifera* vegetative and reproductive organs (*e.g.* branches and flowers) are poorly known. Another issue faced on *Copaifera* exploitation is the challenging identification of its species due to low level of taxonomic description of reproductive organs, the difficulty to obtain fertile specimens and the complexity of the floral structure.[21]

Briefly, *Copaifera* exploitation can be deeply improved with untargeted metabolomics consolidation as a reliable and reproducible method to address such issues involving natural products utilization. Therefore, this work aims to study the viability of high performance methods in the untargeted metabolomic analysis of *Copaifera* species medium polarity extracts fewly explored vegetative organs of this genus, such as seeds and barks, in order to improve the chemical knowledge of this genus, identify chemical markers, and provide insights on untargeted metabolomics applicability.

Results and discussion

Untargeted metabolomic methods showed to be a reliable tool to biodiversity discovery. As most of the *Copaifera* species have few chemical description of their non-volatile composition, the identification of detected metabolites was firstly achieved. This phytochemical screening aided to elucidate the major distinctions and chemical markers among the different extracts (Table 1). The applied method enabled the identification of 29 specialized metabolites of several chemical classes, recognized by their bioactivity, such as galloylquinic acids, flavones, flavanones, flavonols, kaurene diterpenoids, triterpenoids, condensed tannins and xanthones (Table 1). Despite *C. multijuga* and *C. langsdorfii* are well exploited, other species such *C. lucens* and *C. venezuelana* are poorly known.[19,20,22] In fact, there is no previous reports regarding the chemical composition or the bioactivity of *C. venezuelana* extracts, a typical *Copaifera* species endemic from the most attacked part of the Amazon region, the south region known as "Arc of Fire".

Most of identified compounds were flavonoids derived from phenylpropanoid pathway and were present in all evaluated plant organs. Thus, it can be assumed that this biosynthetic pathway presents a great relevance in the construction of *Copaifera* chemical profile. Organs with fewer occurrence of flavonoids were seed and flowers. Their nearly absence in reproductive organs is peculiar, as flavonoids play a role as antioxidant and insect protection within plants, a fundamental tool to enable species propagation.

Galloylquinic acids derivates had already been described within *C. langsdorfii*[23] and *C. multijuga*15 polar extracts and are currently acknowledged as are remarkable gastroprotective and cytotoxic agents.[23] Recently, this chemical class had been indicated as characteristic compound on *Copaifera* leaves.[23–25] Although our data did not identify galloylquinic acids in *C. langsdorfii*[23] and *C. multijuga*, it is the first report of this chemical class within *C. lucens*.

Condensed tannins were only detected within bark extracts (Table 1). Their exploitation within *Copaifera* is recent, with the first description of their presence by Pereira and coworkers in 2018.[26,27] This chemical class had been previously reported as a class of compounds with antioxidant and antineoplastic properties. Their presence observed exclusively in bark extracts can bring insights of

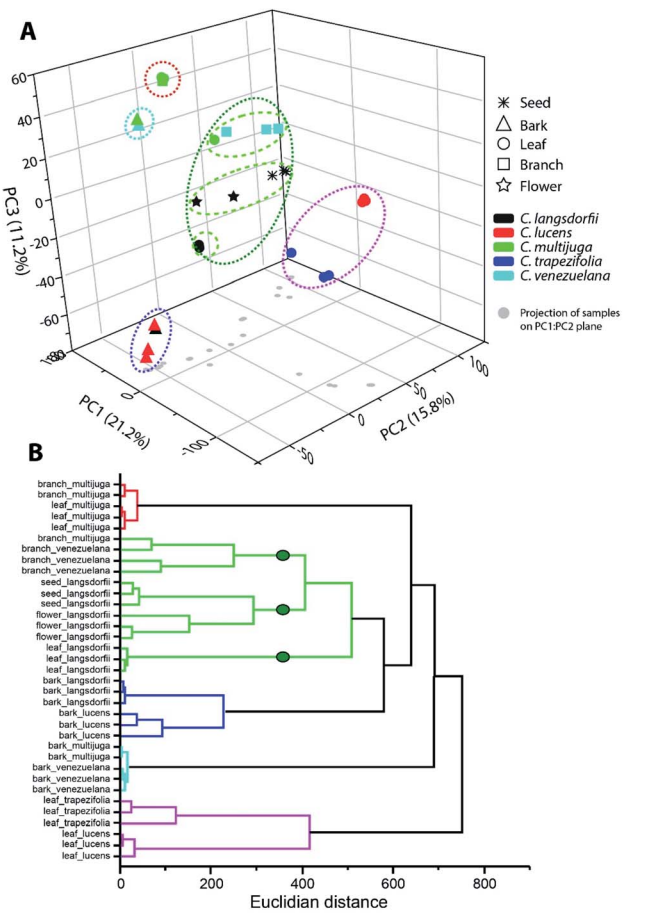

Fig. 1 Multivariate analysis of ethyl acetate fractions by (A) principal component analysis and (B) hierarchical clustering analysis.

new pharmacological approaches to the exploitation of *Copaifera* genus chemophenetic characterization. Also, it can be used to authentication and identification of *Copaifera* bioproducts.

Discrimination of *Copaifera* species and organs

Multivariate analysis was performed to elucidate similarity pattern and exclusive features characteristic of each type of sample, considering plant organs and species differentiation. Within ethyl acetate fractions, samples were clustered in five groups (Fig. 1). Along the groups, two of them (blue and cyan) were composed by bark samples; one group represented the leaves samples (pink group); one represented *C. multijuga* samples (red group); and the last one characterized the *C. langsdorfii* samples.

The blue and cyan group were mainly characterized by the exclusive detection of condensed tannins (**12**, **13** and **22**). Previous description of condensed tannins with copaiba were only within the barks of *C. multijuga*.[15,26] The presence of compounds **13** with *C. lucens* and *C. venezuelana* gives preliminary evidences of their antioxidative activity and an antineoplastic effect against Ehrlich ascites carcinoma, since this diterpene have already being reported as one of the responsible metabolites for *C. multijuga* extracts bioactivity.[18] The occurrence of condensed tannins only in barks indicate such compounds as chemical markers of this organ within the genus and can be considerate to compose a set of chemical parameters to evaluate quality control of herbal products or even identification of the *Copaifera* genus.[7,20,27]

Blue and cyan group also emphasized a relationship among species. The blue group was composed by *C. langsdorfii* and *C. lucens* barks and had as its major characteristic the high relative concentration of ursolic acid (relative concentration higher 60%). Meanwhile, cyan group was composed by *C. multijuga* and *C. venezuelana* barks and characterized by the occurrence of epiafzelechin-(4β,8)-catechin as its base peak. Clustering pattern of those pair of species correlates barks distinction composition with adaptative mechanism related to a specific biome. This hypothesis is possible since *C. multijuga* and *C. venezuelana* are endemic species of rainforest while *C. langsdorfii* and *C. lucens* are typical from Atlantic Forest. The discrimination of geographical origin of botanic samples by chemical markers had already been performed in other genera[28,29] such as described by Cao *et al.* (2021),[30] in which **11** chemical compounds were able to distinguish samples of *Rhizoma Coptidis* from 3 different locations. It is worth mentioning that despite the chemophenetic similarity in the blue group, *C. lucens* bark samples can be distinguished by its exclusive occurrence of gentisein.

The leaves samples of *C. trapezifolia* and *C. lucens* were also clustered in a single group represented in the pink color (Fig. 1). This group evidence the chemical variability among vegetative organs and emphasizes the existence of distinct adaptative processes to each biome as *C. trapezifolia* is also an endemic species of Atlantic Forest. Depending on the environmental stress to which a botanic species is submitted, it will adapt to produce specific sets of specialized metabolites that can increase its survivor. For instance, flavonoids and terpenes accumulation by plants can be increased or decreased under heat or cold stress.[31]

The pink group was mainly characterized by the *m/z* signal 293.176 (retention time of 7.39 min.). Despite the absence of a suggestive identification for this peaks, they were the major compound in pink group samples, distinguishing them of other evaluated samples. Within this group of leaves samples, the *m/z* 251.092 (retention time 5.57) was also of exclusive occurrence among the sampling universe though its relative concentration in *C. trapezifolia* leaves samples was lower than 15%, while it was the base peak of *C. lucens* leaves. The *m/z* 251.092 can be further elucidated to stablish its identity and reliability as a *C. lucens* chemical marker.

Loadings analysis of PCA also emphasizes that quercitrin was detected only in leaves samples, with exception to *C. multijuga*.

Along ethyl acetate samples it is important to highlight the two groups which were able to characterize species (red and green group (Fig. 1)), as species characterization regardless of the plant organ can aid chemophenetic distinction and development of species identification keys. The red group clustered *C. multijulga* leaves and branches samples due to their base peak as 6β,7β-dihydroxykaurenoic acid (**4**). This compound was nearly absent within the other samples, with a relative peak area lower than 9% demonstrating its potential as one chemical marker for *C. multijuga*. The green group was the largest and mainly represented the samples of *C. langsdorfii*. This group presented the formation of three subgroups separating: *C. langsdorfii* leaves; *C. langsdorfii* reproductive organs (seeds and flower); and *C. venezuelana* branches samples. The leaves of *C. langsdorfii* were distinct by their higher concentration of flavonoids, with its major compounds as quercitrin, cirsimaritin and 3,7-di-*O*-methylquercetin. Meanwhile, *C. langsdorfii* reproductive organs were characterized by the unidentified *m/z* 543.442 and 293.212, as *C. venezuelana* branches were characterized by *m/z* 165.054 and 121.065. Despite the heterogeneity of green group, it was still possible to identify chemical markers to distinguish it type of sample by its classification as vegetative or reproductive organ.

Multivariate analysis of methanolic fraction (Fig. 2) corroborates data obtained with ethyl acetate fractions (Fig. 1). Though, distinct *m/z* were emphasized and can be used to distinguish samples. This brings forth additional markers for the chemophenetics of woody *Copaifera* parts and *C. lucens* and *C. langsdorfii*. Within methanol analysis (Fig. 2A), the leaves of *C. lucens* and *C. multijuga* were clustered based in the detection of quinic acid on them. Although it is a well-reported metabolite within *Copaifera* genus,[19,25,32] its application together with the other discriminate *m/z* signals described in our data demonstrate its capacity to distinguish plants organs. In addition, leaves of both species have different major peaks. Whilst *C. lucens* base peak was *m/z* 647.088, *C. multijuga* major detected signal was *m/z* 689.135 and quinic acid.

Methanolic fraction also enable to distinguish the barks of *C. lucens* of the *C. langsdorfii* (Fig. 2A, blue group). Characterizing signals of *C. lucens* barks were *m/z* 542.260 and 525.270. Furthermore, *C. venezuelana* and *C. multijuga* were clustered together though several of its major compounds, which were not detected in the other samples (Fig. 2A), such as the putative identified signal 25, a flavonol with a β-glucosyl-(1 → 2)-β-

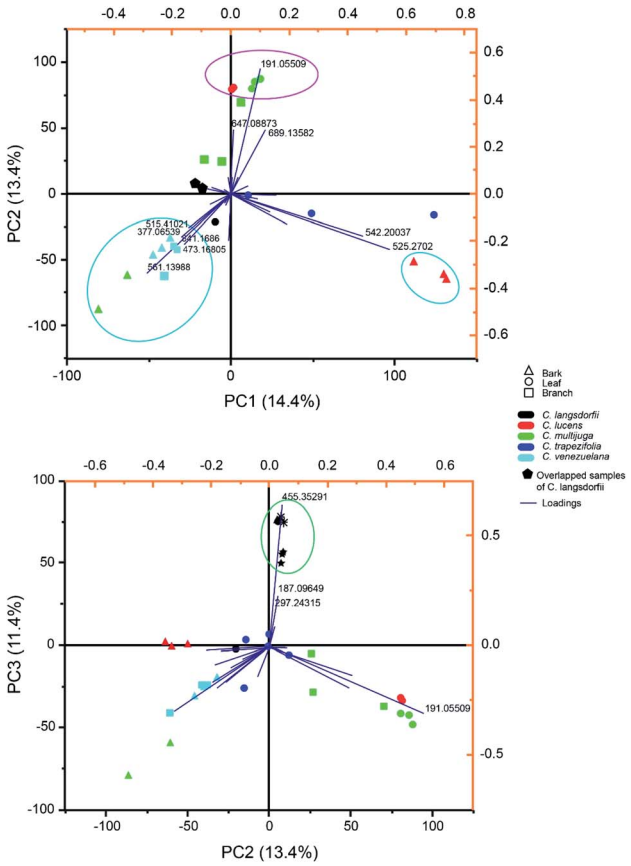

Fig. 2 Principal component analysis of methanolic fraction biploted as (A) principal component (PC) 1 *versus* PC 2 and (B) PC 2 *versus* PC 3.

glucosyl residue at position 3 by a glycosidic linkage. In the case of *C. Langsdorfii*, its samples still presented themself in a consice cluster (Fig. 2B), being discriminated from other samples by ursolic acid, as previosly mentioned and the unindentified *m/z* 187.096 and 297.243.

Experimental

Botanic samples

Botanic samples of five *Copaifera* species were collected at Brazilian Amazon rainforest (Amazonas state) and Atlantic forest (Rio de Janeiro and São Paulo state, Brazil) (Table 2). Voucher of each specimen were deposited at the National Institute of Amazon Researches (INPA) herbarium and Rio de Janeiro Botanical Garden Herbarium (RB). All samples collected (Table 2) were cleaned to remove dust and small insects, air-dried and grinded in a knife mill (Willey Mill SP-32, SPLabor). After sample pretreatment they were stored frozen (−4 °C) until extraction procedures. Botanic material sampling was authorized and registered on the National System of Genetic Resource Management and Associated Traditional Knowledge (SISGEN) system (Table 2) in accordance with Brazilian legislation regarding biodiversity scientific exploitation.

Extraction was conducted by exhaustive maceration in two cycles of 48 hours at room temperature. Each extraction cycle was performed with a distinct solvent, firstly ethyl acetate followed by methanol. The extractor solvent was used in a proportion of 1 : 10 (g : mL) between sample and solvent. After extraction, the obtained extracts were dried in a rotary evaporator under reduced pressure and used in further chemical analysis. Solvent choice was based on previous reports of bioactive extracts and non-volatile chemical fingerprint of copaiba, which suggest the presence of flavonoids, terpenoids and galloylquinic acid derivates.

Chemical profiling

Chemical analyses were performed in an Ultra High-Performance Liquid Chromatographer Dionex Ultimate 3000 UHPLC (Thermofisher Scientific, Bremen, Germany) coupled to High Resolution Mass Spectrometer Q-Exactive (Thermofisher Scientific, Bremen, Germany) (UHPLC-HRMS). Both extracts types were degreased with hexane and prepared with methanol at 1.0 mg mL^{-1}. Next, samples were filtered at a 0.45 μm PTFE w/GMF membranes (Whatman, Little Chalfont, United Kingdom). All samples were prepared and analyzed as triplicates.

Chromatographic analyses were accomplished in an exploratory gradient elution in a Syncronis C18 (2.1 × 50 mm, 100 Å – Thermofisher Scientific, Waltham, United States) and a mobile phase composed by solvent A (formic acid 0.1%: ammonium formiate 5 mM) and solvent B (methanol). The elution gradient ranged from 0.0 to 100% of solvent B in 9 minutes; 100% of solvent B for 1 minute; and a gradient of 100 to 0.0% of solvent B in 1 minute. The mass spectrometer was equipped with an electrospray ionization source operating in negative mode for

Table 2 *Copaifera* samples description

Species	Sampling site	Geographic coordinates	Voucher number	SISGEN number	Collected samples
C. multijuga hayne	Manaus/AM	3° 0′ 27.00″ S 59° 56′ 22.92″ W	INPA82418	A09F694	Leaf, bark and branch,
C. venezuelana harms & Pittier	Coari/AM	4°05′37.3″S 63°09′04.5″W	INPA229668	A9CDF0F	Bark and branch
C. langsdorfii Desf.	Campinas/SP	22°49′02.9″S 47°04′12.1″W	RB344931	A3AE56C	Leaf, bark, branch, seed and flower
C. trapezifolia hayne	São miguel arcanjo/SP	24°03′18.2″S 47°59′40.5″W	RB142240	ACDC561	Leaf
C. lucens Dwier	Rio de janeiro/RJ	22°58′03.1″S 43°13′29.0″W	RB301745	A10C77E	Leaf and bark

data acquisition in MS1 and MS2 at the range of 100 to 900 *m/z*. MS2 experiments were performed in a data dependent acquisition mode. Operational parameters of UHPLC and HRMS are available at ESI (Tables S1 and S2).†

Data analysis

Peak detection and integration were performed on MZmine 2 v2.53 though a workflow which included baseline correction, mass detection, deconvolution, deisotoping and alignment of each spectra through the algorithms asymmetric least square baseline correction, exact mass, baseline cut-off, isotopic peak grouper and join aligner.[22,33,34] MZmine 2 configurations on each step are available at ESI (Table S3).†

The identification of compounds present on the detected peaks were performed though analysis of the MS^2 fragmentation spectra by similarity with the databases of Keggs,[22,33–37] and with the Feature-based Molecular Networking (FBMN) workflow on the Global Natural Products Social Molecular Networking (GNPS). Such workflows are recommended protocols to identify compounds on untargeted metabolomic methods.[34–39] In GNPS-FBMN the fragmentation spectra of the sample is compared with the database and the similarity between them is evaluated based in the ppm error of the precursor ions, presence of fragments, and their relative intensity. The similarity is measure as a cosine score, that range from 0 (completely different) to 1 (identical). Depending on the resolution of the mass spectrometer, a cosine score above 0.7 can produce results with nearly 90% accuracy.[20] For FMBN-GNPS identification of compounds, samples fragmentation spectra were only compared with data from the database which have a maximum *m/z* difference of ±17 Da among samples and database precursor ion and fragment ions. A positive identification was performed only with samples and database spectra presented at least 3 identical fragment ions and a cosine score above 0.7. In addition, for positive identifications, mass tolerance for the precursor and fragment ions was set as 0.2 Da.[34–38]

Chemometric analysis was performed on OriginPro 2017 by hierarchical clustering analysis (HCA) with Ward's method and Euclidian distance and Principal Component Analysis (PCA), with confidence levels of 95% ($\alpha = 0.05$). In this analysis, each detected peak was considered as a variable in which their integrated relative areas were the response variable.

Conclusions

Untargeted metabolomic approach evidences the ability to differentiate *Copaifera* species even if morphological features of the specimens cannot be observed, using non-volatile extracts. The UHPLC-HRMS/MS method provided reliable and distinct chemical fingerprint for five *Copaifera* species, enabling the identification of 29 substances within flavonoids, diterpenes and galloylquinic acids. The identification of metabolites provided insight of *Copaifera* chemodiversity, especially regarding *C. trapezifolia* and *C. lucens* chemical composition. Both species are poorly scientifically explored and presented in our data galloylquinic acids and flavonoids, which are known by their remarkable bioactivities. In addition, 19 chemical markers were highlighted by chemometric analysis which enabled the characterization of samples by their geographic distribution and efficiently distinct each species and their vegetative organs. Flavonoids, condensed tannins, galloylquinic acids and kaurenoic acid derivates demonstrated great chemophenetic relevance by untargeted approaches. Untargeted metabolomic showed to be a feasible approach to elucidate several chemical issues of complex botanic groups, by such as *Copaifera*, generating massive amount of chemical data with very few sampling size.

Author contributions

Ananda S. Antonio: Formal analysis, Investigation, Data Curation, Writing – Original Draft, Visualization. Davi S. Oliveira and Gustavo R. C. dos Santos: Investigation. Henrique M. G. Pereira: Resources, Writing – Review & Editing, Supervision. Larissa S. M. Wiedemann and Valdir F. Veiga-Junior: Conceptualization, Methodology, Resources, Writing – Review & Editing, Visualization, Supervision, Project administration, Funding acquisition.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. The authors would like to acknowledge the financial support of FAPERJ and the National Council for Scientific and Technological Development (CNPq).

References

1 L. M. Ricardo, B. M. Dias, F. L. B. Mügge, V. V. Leite and M. G. L. Brandão, *J. Ethnopharmacol.*, 2018, **219**, 319–336.
2 P. Ibarra-Torres, E. Valadez-Moctezuma, M. Pérez-Grajales, J. Rodríguez-Campos and M. E. Jaramillo-Flores, *Sci. Hortic.*, 2014, **181**, 137–146.
3 M. Bakha, N. El Mtili, N. Machon, K. Aboukhalid, F. Z. Amchra, A. Khiraoui, M. Gibernau, F. Tomi and C. Al Faiz, *Arabian J. Chem.*, 2020, **13**, 3070–3081.
4 S. Cao, H. Du, B. Tang, C. Xi and Z. Chen, *Microchem. J.*, 2021, **160**, 105685.
5 G. A. B. Canuto, J. L. da Costa, P. L. R. da Cruz, A. R. L. Souza, A. T. Faccio, A. Klassen, K. T. Rodrigues and M. F. M. Tavares, *Quim. Nova*, 2018, **41**, 75–91.
6 H. Gika, C. Virgiliou, G. Theodoridis, R. S. Plumb and I. D. Wilson, *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.*, 2019, **1117**, 136–147.
7 O. Begol, H. G. Gika, G. A. Theodoridis and I. D. Wilson, in *Metabolomic Profiling Methods Mol. Biol.*, ed. G. Theodoridis, H. Gika and I. Wilson, Humana Press, New York, 1st edn, 2018.

8 E. Gorrochategui, J. Jaumot, S. Lacorte and R. Tauler, *Trends Anal. Chem.*, 2016, **82**, 425–442.
9 A. C. Schrimpe-rutledge, S. G. Codreanu, S. D. Sherrod and J. A. Mclean, *J. Am. Soc. Mass Spectrom.*, 2017, **27**, 1897–1905.
10 M. Lv, Y. Tian, Z. Zhang, J. Liang, F. Xu and J. Sun, *RSC Adv.*, 2015, **5**, 15700–15708.
11 F. Van Der Kooy, F. Maltese, H. C. Young, K. K. Hye and R. Verpoorte, *Planta Med.*, 2009, **75**, 763–775.
12 F. de S. Vargas, P. D. O. Almeida and E. S. P. Aranha, *Molecules*, 2015, **20**, 6194–6210.
13 K. F. Vasconcelos, V. F. da Veiga Junior and W. C. Rocha, *Rev. Bras. Farm.*, 2008, **18**, 733–738.
14 J. P. B. Sousa, D. Nanayakkara, A. A. B. Silva and J. K. Bastos, *J. Pharm. Res.*, 2012, **5**, 4103–4107.
15 D. Oliveira, A. S. Antonio, L. Lima, L. Wiedemann and V. Veiga-Junior, *Quim. Nova*, 2019, **43**, 1–31.
16 J. F. Carmo, I. Miranda, T. Quilhó, V. B. Sousa, S. Cardoso, A. M. Carvalho, F. H. D. J. Carmo, J. V. F. Latorraca and H. Pereira, *J. Wood Chem. Technol.*, 2016, **36**, 305–317.
17 F. T. T. Trindade, R. G. Stabeli, A. A. Pereira, V. A. Facundo and A. de A. Silva, *Brazilian J. Pharmacogn.*, 2013, **23**, 464–470.
18 A. P. S. da Cunha, L. Baldissera, D. L. Pereira, L. R. Albiero, L. Castoldi, A. P. Sinhorin and V. D. G. Sinhorin, *Acta Amazonica*, 2019, **49**, 41–47.
19 R. Trindade, J. K. da Silva and W. N. Setzer, *Int. J. Mol. Sci.*, 2018, **19**, 1–33.
20 C. Arruda, J. A. Aldana Mejía, V. P. Ribeiro, C. H. Gambeta Borges, C. H. Gomes Martins, R. C. Sola Veneziani, S. R. Ambrósio and J. K. Bastos, *Biomed. Pharmacother.*, 2019, **109**, 1–20.
21 R. C. V. Martins-da-Silva, J. F. Pereira and H. C. de Lima, *Rodriguesia*, 2008, **59**, 455–476.
22 O. D. Myers, S. J. Sumner, S. Li, S. Barnes and X. Du, *Anal. Chem.*, 2017, **89**, 8696–8703.
23 E. V. S. Motta, M. Lemos, J. C. Costa, V. C. Banderó-Filho, A. Sasse, H. Sheridan and J. K. Bastos, *Chem.-Biol. Interact.*, 2017, **261**, 145–155.
24 R. da Trindade, J. K. da Silva and W. N. Setzer, *Int. J. Mol. Sci.*, 2018, **19**, 1511.
25 M. S. Nogueira, R. A. Furtado and J. K. Bastos, *J. Agric. Food Chem.*, 2015, **63**, 6939–6945.
26 D. L. Pereira, A. P. S. Da cunha, C. R. P. Cardoso, C. Q. Da rocha, W. Vilegas, A. P. Sinhorin and V. D. G. Sinhorin, *Acta Amazonica*, 2018, **48**, 347–357.
27 L. J. Carneiro, T. C. Bianchi, J. M. J. da Silva, L. C. Oliveira, C. H. G. Borges, D. C. Lemes, J. K. Bastos, R. C. S. Veneziani and S. R. Ambrósio, *J. Braz. Chem. Soc.*, 2018, **4**, 729–737.
28 D. T. Burns, L. Tweed and M. J. Walker, *Food Anal. Methods*, 2017, **10**, 2302–2310.
29 A. Shevchuk, L. Jayasinghe and N. Kuhnert, *Food Res. Int.*, 2018, **109**, 387–402.
30 S. Cao, H. Du, B. Tang, C. Xi and Z. Chen, *Microchem. J.*, 2021, **160**, 105686.
31 M. A. Ashraf, M. Iqbal, R. Rasheed, I. Hussain, M. Riaz and M. S. Arif, in *Plant Metabolites and Regulation Under Environmental Stress*, P. Ahmad, M. A. Ahanger, V.P. Singh, D. K. Tripathi, P. Alam and M. N. Alyemeni, Academic Press, London, 1st edn, 2018.
32 D. S. Oliveira, L. S. Lima, A. S. Antonio, L. S. M. Wiedemann and V. F. Veiga-Junior, *Quim. Nova*, 2020, **43**, 72–77.
33 T. Pluskal, S. Castillo, A. Villar-Briones and M. Orešič, *BMC Bioinf.*, 2010, **395**, 1–11.
34 A. T. Aron, E. C. Gentry, K. L. McPhail, L. F. Nothias, M. Nothias-Esposito, A. Bouslimani, D. Petras, J. M. Gauglitz, N. Sikora, F. Vargas, J. J. J. van der Hooft, M. Ernst, K. Bin Kang, C. M. Aceves, A. M. Caraballo-Rodríguez, I. Koester, K. C. Weldon, S. Bertrand, C. Roullier, K. Sun, R. M. Tehan, C. A. Boya P, M. H. Christian, M. Gutiérrez, A. M. Ulloa, J. A. Tejeda Mora, R. Mojica-Flores, J. Lakey-Beitia, V. Vásquez-Chaves, Y. Zhang, A. I. Calderón, N. Tayler, R. A. Keyzers, F. Tugizimana, N. Ndlovu, A. A. Aksenov, A. K. Jarmusch, R. Schmid, A. W. Truman, N. Bandeira, M. Wang and P. C. Dorrestein, *Nat. Protoc.*, 2020, **15**, 1954–1991.
35 L. F. Nothias, D. Petras, R. Schmid, K. Dührkop, J. Rainer, A. Sarvepalli, I. Protsyuk, M. Ernst, H. Tsugawa, M. Fleischauer, F. Aicheler, A. Aksenov, O. Alka, P.-M. Allard, A. Barsch, X. Cachet, M. Caraballo, R. R. Da Silva, T. Dang, N. Garg, J. M. Gauglitz, A. Gurevich, G. Isaac, A. K. Jarmusch, Z. Kameník, K. Bin Kang, N. Kessler, I. Koester, A. Korf, A. Le Gouellec, M. Ludwig, M. H. Christian, L.-I. McCall, J. McSayles, S. W. Meyer, H. Mohimani, M. Morsy, O. Moyne, S. Neumann, H. Neuweger, N. H. Nguyen, M. Nothias-Esposito, J. Paolini, V. V. Phelan, T. Pluskal, R. A. Quinn, S. Rogers, B. Shrestha, A. Tripathi, J. J. J. van der Hooft, F. Vargas, K. C. Weldon, M. Witting, H. Yang, Z. Zhang, F. Zubeil, O. Kohlbacher, S. Böcker, T. Alexandrov, N. Bandeira, M. Wang and P. C. Dorrestei, *Nat. Methods*, 2020, **17**, 905.
36 H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito and T. Nishioka, *J. Mass Spectrom.*, 2010, **45**, 703–714.
37 M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan, C. Porto, A. Bouslimani, A. V. Melnik, M. J. Meehan, W. T. Liu, M. Crüsemann, P. D. Boudreau, E. Esquenazi, M. Sandoval-Calderón, R. D. Kersten, L. A. Pace, R. A. Quinn, K. R. Duncan, C. C. Hsu, D. J. Floros, R. G. Gavilan, K. Kleigrewe, T. Northen, R. J. Dutton, D. Parrot, E. E. Carlson, B. Aigle, C. F. Michelsen, L. Jelsbak, C. Sohlenkamp, P. Pevzner, A. Edlund, J. McLean, J. Piel, B. T. Murphy, L. Gerwick, C. C. Liaw, Y. L. Yang, H. U. Humpf, M. Maansson, R. A. Keyzers, A. C. Sims, A. R. Johnson, A. M. Sidebottom, B. E. Sedio, A. Klitgaard, C. B. Larson, C. A. P. Boya, D. Torres-Mendoza, D. J. Gonzalez, D. B. Silva, L. M. Marques, D. P. Demarque, E. Pociute, E. C. O'Neill,

E. Briand, E. J. N. Helfrich, E. A. Granatosky, E. Glukhov, F. Ryffel, H. Houson, H. Mohimani, J. J. Kharbush, Y. Zeng, J. A. Vorholt, K. L. Kurita, P. Charusanti, K. L. McPhail, K. F. Nielsen, L. Vuong, M. Elfeki, M. F. Traxler, N. Engene, N. Koyama, O. B. Vining, R. Baric, R. R. Silva, S. J. Mascuch, S. Tomasi, S. Jenkins, V. Macherla, T. Hoffman, V. Agarwal, P. G. Williams, J. Dai, R. Neupane, J. Gurr, A. M. C. Rodríguez, A. Lamsa, C. Zhang, K. Dorrestein, B. M. Duggan, J. Almaliti, P. M. Allard, P. Phapale, L. F. Nothias, T. Alexandrov, M. Litaudon, J. L. Wolfender, J. E. Kyle, T. O. Metz, T. Peryea, D. T. Nguyen, D. VanLeer, P. Shinn, A. Jadhav, R. Müller, K. M. Waters, W. Shi, X. Liu, L. Zhang, R. Knight, P. R. Jensen, B. Palsson, K. Pogliano, R. G. Linington, M. Gutiérrez, N. P. Lopes, W. H. Gerwick, B. S. Moore, P. C. Dorrestein and N. Bandeira, *Nat. Biotechnol.*, 2016, **34**, 828–837.

38 H. Mohimani, A. Gurevich, A. Shlemov, A. Mikheenko, A. Korobeynikov, L. Cao, E. Shcherbin, L. F. Nothias, P. C. Dorrestein and P. A. Pevzner, *Nat. Commun.*, 2018, **9**, 1–12.

39 K. Scheubbert, F. Hufsky, D. Petras, M. Wang, L. Nothias, K. Duhrkop, N. Bandeira, P. C. Dorrestein and S. Bocker, *Nat. Commun.*, 2017, **8**, 1494.