Integrating chemotaxonomic-based metabolomics data with DNA barcoding for plant identification: A case study on south-east African Erythroxylaceae species.

P.S.F. Alberts [a] *, J.J.M. Meyer [a]

[a] *Department of Plant and Soil Sciences, University of Pretoria, Pretoria, 0002, South Africa*

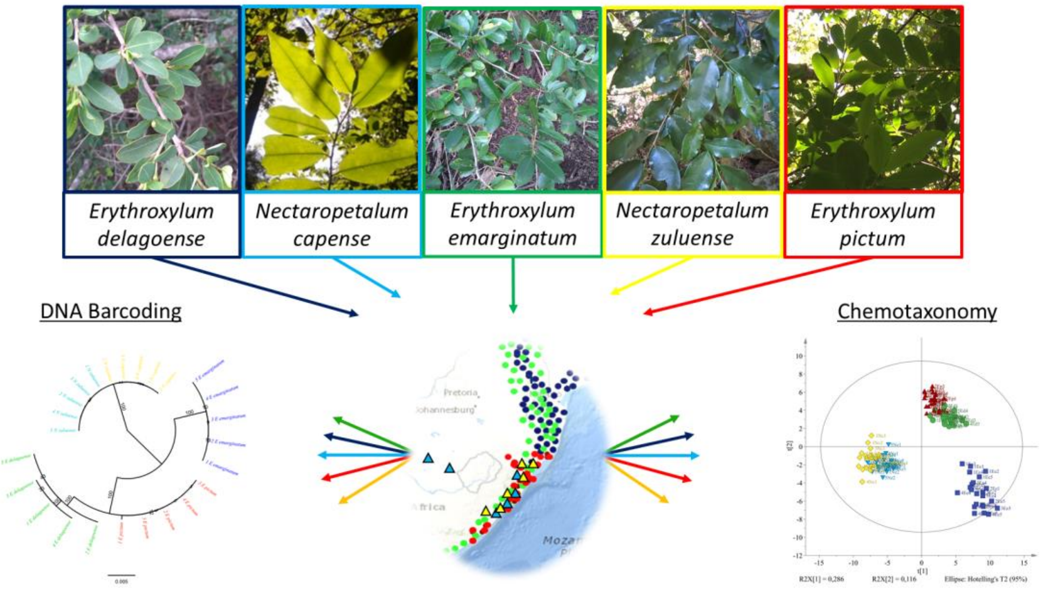
*Author for correspondence:

P.S.F. Alberts

Tel: +27 12 420 2224

E-mail: sewes.alberts@up.ac.za

Graphical abstract



Highlights

- The identification and classification of closely related taxa are described.
- Identifying characteristics were compared between chemotaxonomy and DNA barcoding.
- The identification accuracy increased from morphology to DNA barcoding to chemotaxonomy.
- Phylogenetic clades correlate well with plant metabolite clustering.
- The financial cost implications of the different plant identification methodologies are described.

Abstract

Plants have been used as medicines for millennia and are major contributors to developed western pharmacopoeia. The *Erythroxylum* and *Nectaropetalum* genera belong to the Erythroxylaceae (coca) family, with select species capable of producing highly valued 'blockbuster' medicinal compounds including, amongst others atropine, cocaine, scopolamine, and tigloidine. *Erythroxylum delagoense, E. emarginatum, E. pictum, N. capense* and *N. zuluense* are indigenous to the south-east tropical regions of Africa. The morphological similarity between these taxa make identification to species-level troublesome and often unreliable, indicating a need for alternative identification methods. This study aimed to compare gas chromatography-mass spectrometry (GC-MS)- and nuclear magnetic resonance (NMR)-based metabolomics analyses with DNA barcoding to evaluate the identifying characteristics of these coca species. The results emphasise the importance of integrating chemotaxonomy and DNA barcoding techniques in plant identification. In this sample of the Erythroxylaceae, the differentiating identification accuracy was shown to increase from morphology to DNA barcoding to chemotaxonomy. This study further highlights the strengths and weaknesses of various plant identification strategies, as well as providing a developing model for more accurate and reliable species-level identification of plants. The findings from this case study could aid in the identification and classification of other closely related taxa.

Key words: *Erythroxylum, Nectaropetalum;* GC-MS*;* NMR; Metabolomics; DNA barcoding.

1. Introduction

The Erythroxylaceae and Solanaceae families contain some of the best-known medicinal plants with select species producing a variety of, or precursors to, highly valued, pharmaceutically active compounds including atropine, hyoscyamine, scopolamine, tiotropium bromide, ipratropium bromide and cocaine (Grynkiewicz and Gadzikowska, 2008; Hasan, 2012). Five species within the Erythroxylaceae family, distributed along the south-east coast and tropical regions of Africa, were the focus of this study. These include *Erythroxylum delagoense* Schinz, *E. emarginatum* Thonn., *E. pictum* E.Mey. ex Harv. & Sond., *Nectaropetalum capense* (Bolus) Stapf & Boodle and *N. zuluense* (Schönland) Corbishley. The traditional uses of these species indicate pharmaceutical potential (Corrigan et al., 2011; de Wet, 2011; Nishiyama et al., 2007), however, as a result of shared morphological characteristics, accurate species-level identification of non-flowering material is difficult, as is the case with other species within the Erythroxylaceae (White et al., 2019).

Morphological evaluation of the two genera in South Africa indicates that species of *Erythroxylum* display more phenotypic differences than the species of *Nectaropetalum* (Coates-Palgrave, 2002; Boon, 2010; Van Wyk & van Wyk, 2013). The distinguishing characteristics between the two *Nectaropetalum* species lie within the leaf morphology and flower structure. *Nectaropetalum capense* has lighter and more transparent leaf veins compared to *N. zuluense*, while the flowers of *N. capense* have shorter stalks (1.5 mm) and no nectar pocket at the base of the petals. This is in contrast to the long flower stalks of *N. zuluense* (10 mm) and the presence of a nectar pocket at the petal base (Coates-Palgrave, 2002; Boon, 2010; Van Wyk & van Wyk, 2013). The *Erythroxylum* species are vegetatively more distinct, with differences in leaf size and colour (Boon, 2010; Coates-Palgrave, 2002). Confusion and misidentification can still occur when sampling in the field, as the morphological traits are not pronounced, while further being obscured by phenotypic plasticity (Fusco and Minelli, 2010; Gratani, 2014; Mantuano et al., 2006).

These difficulties may be overcome by incorporating datasets from alternative identification techniques such as chemotaxonomy and DNA barcoding (Govaerts, 2001; Mishra et al., 2016; Mora et al., 2011; Scotland and Wortley, 2003). The effectiveness of chemotaxonomy as an identification and authentication tool has been shown (Afzan et al., 2019; Ahmad et al., 2010, 2009; Endara et al., 2018; Heyman and Meyer, 2012; Mishra et al., 2016), with select studies investigating differentiating biomarkers related to specific *Erythroxylum* species (Johnson et al., 1998; Zanolari et al., 2005). However, to our knowledge, there have been no chemometric studies done on any of the south-east African coca species. Chemotaxonomy utilises a variety of technologies, with nuclear magnetic resonance (NMR) and mass spectrometry (MS) related techniques generating accurate, reliable and reproducible data as well as being readily used in metabolomic studies (Afzan et al., 2019; Alonso et al., 2015; Endara et al., 2018; Jorge et al., 2016).

Furthermore, the use of standardised predefined DNA barcodes in plant identification is well established, enabling scientists not specialised in taxonomy to identify and distinguish between different taxa, and unknown plant samples lacking in morphological traits (Afzan et al., 2019; CBOL Plant Working Group, 2009; Dev et al., 2014; Kress et al., 2009, 2005; Mishra et al., 2016; Yu et al., 2021). The use of multiple barcode regions is strongly advised for increased identification accuracy, as factors such as polyploidy, hybridisation and generation periods of plants could affect the sequence similarity between species (Smith and Donoghue, 2008; Spooner, 2009). These factors contribute to the difficulty of assigning a universal barcode for the identification of plants.

The most readily used and recommended DNA barcoding regions are the coding loci *rbcL* (ribulose-1,5-bisphosphate carboxylase/ oxygenase large subunit), *matK* (maturase K) and *trnH-psbA* (intergenic spacer region), found in the chloroplast DNA of plants (Bolson et al., 2015; CBOL Plant Working Group, 2009; Kress et al., 2009). These gene regions, when used in combination, have the highest universality and discrimination power, lowest sequencing cost and highest quality (CBOL Plant Working Group, 2009; Kress et al., 2009). Using these loci, identification to genus level can be achieved, in most cases, regardless of the growth form or developmental stage, although this is dependent on the availability of a robust sequence reference library (de Lima et al., 2018; de Vere et al., 2012; Mishra et al., 2016; Seberg and Petersen, 2009).

In this study we report on the integrated use of gas chromatography-mass spectrometry (GC-MS) and NMR-based chemotaxonomy, with DNA barcoding, to assess their combined and separate identification capabilities. The data from this case study provides proof-of-concept and could be utilised in future development of more robust plant identification systems.

2. Materials and Methods

2.1. Plant material collection and processing

Healthy leaf material of varying ages was collected during autumn (2016) from three *Erythroxylum* and two *Nectaropetalum* species found in nature reserves along the east coast of South Africa (Table S. 1), including Enseleni Nature Reserve (*E. delagoense* Schinz), Mpenjati Nature Reserve (*E. emarginatum* Thonn.) and Umtamvuna Nature Reserve (*E. pictum* E.Mey. ex Harv. & Sond., *N. capense* (Bolus) Stapf & Boodle, and *N. zuluense* (Schönland) Corbishley).

Biological replicates were prepared by sampling leaf material from five different trees within each species. The average distances between the individual trees of a species were five to ten meters. Leaf material used in the chemotaxonomy study, were sampled from each of the four cardinal directions, as well as the top and bottom of the trees. The leaves of each respective tree (± 128 g) were mixed and divided into five technical replicates, thus ensuring representation of the entire tree and homogenising the chemical variation within each specimen as far as possible.

The collection for the DNA barcoding study entailed sampling 10 to 20 healthy young leaves from each of the same five individual trees per species. To account for the possible genetic variation within an individual, the leaf material was harvested from different sub-branches of the main trunk of each tree. Thus, providing a representative DNA sample from which the barcodes could be extracted and assessed.

Preservation of the leaf material used in both the chemotaxonomy and DNA barcoding studies during the collection period, was done by placing the leaves of each sample in separate, appropriately labelled paper bags. Each paper bag was placed inside a plastic zip-lock bag containing silica gel beads (± 8 g) to dehydrate the leaves whilst limiting compound and DNA breakdown (Chase and Hills, 1991; Wilkie et al., 2013). The samples were kept on blocks of dry ice during transportation. Flash freezing with liquid nitrogen was not feasible as a result of the travel time and remote localities of the trees.

Upon arrival at the laboratory, plant material used in the chemotaxonomy study was lyophilised using a benchtop freeze-dryer (VirTis, United Scientific, USA) to ensure full dehydration and to limit the breakdown of compounds in the leaf samples (Abascal et al., 2005; Bhatta et al., 2020). The dried leaf material was thereafter stored at 4°C until crude plant extracts could be prepared. The material used in the DNA barcoding study was stored at -80ºC until DNA could be extracted.

Representative voucher specimens of all the individual trees studied have been deposited in the H.G.W.J. Schweickerdt Herbarium at the University of Pretoria (PRU) (see Table S. 1). The plant material was collected in accordance with the regulations of the Department of Environmental Affairs, the Ezemvelo KZN Wildlife conservation agency and the South African National Biodiversity Institute (SANBI) (Permit numbers OP567/2016 and BABS/000515N).

2.2.1. Chemotaxonomy: Extract preparation

Pressurised methanolic crude leaf extracts were prepared as described in a previous study (Alberts et al., 2018).

2.2.2. Chemotaxonomy: ^{1}H-NMR analysis

The NMR analysis was conducted using 12 mg vacuum dried (GeneVac EZ-2 plus; SP Industries Inc., Warminster, England) crude methanolic extract, dissolved in 800 µl (15 mg ml^{-1}) deuterated methanol (CD_3OD) by sonication for 5 min. Proton (^{1}H) NMR was performed on each of the respective samples, using 700 µl of this solution, on a 200 MHz NMR spectrometer (Varian, Palo Alto, California). The standard 1D spectra were acquired with the total transients set to 256. This included 11976 data points at a spectral width of 3003 Hz (-2 to 12 ppm), a three-second relaxation delay, and the acquisition time for each transient scan set to two seconds. The magnet shimming was done manually for optimal and consistent spectral resolution. Each sample was analysed directly after preparation and the run order was randomised by dividing the samples into five batches, based on replicate number i.e., batch one contained the first replicate of each tree, batch two the second replicate etc.

MestReNova (Mnova) version 14.2 analytical software (Mestrelab Research S.L., 2020) was used to process the NMR data, which included referencing the spectra (residual CH_3OH; δ3.310 ppm), baseline correction (Whittacker smoother), automatic phase correction, and normalisation based on the residual CH_3OH signal (δ3.310 ppm to intensity = 100) in the deuterated CD_3OD. The respective spectra were binned (0.04 ppm per bin) and exported to the soft independent modelling by class analogy (SIMCA-P) analytical software version 14.1 (Umetrics, Umeå, Sweden) for associated multivariate data analysis (MVDA). The specific bin sizes were chosen as they generated a higher statistically significant interpretation of the data as compared to the 0.10, 0.02 and 0.08 binned data (data not shown). The peak signals at chemical shifts between δ3.27–

δ3.34 ppm and δ4.80–δ5.00 ppm was removed before SIMCA-P processing as these regions contained residual solvent peaks.

2.2.3. Chemotaxonomy: GC-MS analysis

The GC-MS samples contained 1 mg ml^{-1} underivatised crude extract in distilled methanol for untargeted metabolomic analysis. Samples were analysed on a Shimadzu GC-MS-QP2010 (Shimadzu Corporation, Japan) with an electric current of 70 eV. The compounds were separated using an Rtx-5MS column (29.3 m x 250 µm x 0.25 µm i.d.; 0.25 µm df) with helium as the carrier gas. Splitless injections of 1 µl were performed, with the column flow set to linear velocity. The injector and interface temperatures were set at 250°C, with the ion source temperature set to 200°C. The oven temperature program was set to an initial 50°C and held for 2 min, thereafter, increased at a rate of 10°C min^{-1} to 300°C, which was held for 5 min, bringing the total run time to 30 min per sample. The mass hertz (*m/z*) detection range was set from 50 to 600 *m/z* with a scan speed of 2000 aum s^{-1}. Blank methanolic samples were introduced after every five samples for quality control purposes. The samples were prepared and directly loaded onto an autosampler. The run order was randomised as described in the ^{1}H-NMR analysis section.

The GC-MS chromatograms were analysed using the Shimadzu GC-MS LabSolutions software version 4.2 (Shimadzu Corporation, Japan). The total ion current (TIC) chromatograms were exported as *.csv (MS-DOS) files from the Shimadzu GC-MS LabSolutions software and imported into Mnova, where baseline correction (Whittacker smoother), spectral alignment (Rt 17.09 – neophytadine), normalising and binning was done. The processed chromatograms were then exported to SIMCA-P for further MVDA.

The binning size of Rt 0.04 was chosen as it showed higher statistical significance and resolution of the compound peaks (data not shown).

2.2.4 Chemotaxonomy: GC-MS-based biomarkers

Biomarkers were identified for each species by comparing the GC-MS data to the National Institute of Standards and Technology (NIST 11) software database (version 2.2, Agilent Technologies, USA). Biomarker identification was done by generating a SIMCA-P orthogonal projection to latent structures discriminant analysis (OPLS-DA) (Trygg and Wold, 2002; Wiklund et al., 2008) plot with two classes. Class one containing the data from a specific species group (e.g. *E. delagoense*) and class two containing all the data from the remaining four species. Each supervised OPLS-DA plot was found to be significant based on Permutation (100 permutations) and cross validation (CV)-ANOVA (p-value < 0.05) scores. Following this, an S-Plot and a Contribution Plot corresponding to each OPLS-DA model was generated. The outliers (differentiating variables) on the S-Plot were extracted and identified in the corresponding Contribution Plot, revealing the retention times of the variables which contribute the most in differentiating between the two classes, i.e. variables outside the third standard deviation range of the respective data set. This was then related back to the TIC chromatogram on the Shimadzu Post-run analysis, where the mass fragmentation pattern correlating to the peak of interest (differentiating variable) was extracted and searched on the NIST 11 spectral database for a possible match. A similarity match factor below 85% was not considered for subsequent analysis (Stein, 1999). The biomarkers identified were further validated by visually comparing the mass spectral fragmentation pattern, as well as the *m/z* peak abundance, to the data found in the NIST library (Supplementary Fig. S 1). A graphical summary of this process can be seen in the Results and Discussion section.

2.3.1.DNA barcoding: DNA extractions

The total DNA was extracted using the cetyltrimethylammonium bromide (CTAB) method adapted from Doyle & Dickson (1987), Doyle and Doyle (1987), and Cullings (1992). Quantification of the DNA was done using a NanoDrop™ 2000 spectrometer (Thermo Fisher Scientific, Inc., United States) and agarose gel (Lonza SeaKem® LE Agarose, Switzerland) electrophoresis.

2.3.2. DNA barcoding: DNA barcode sequence generation

Three DNA barcoding regions, *rbcL*, *matK*, and a chloroplast spacer region *trnH-psbA* proposed by the CBOL Plant Working Group was assessed during this study (Bolson et al., 2015; CBOL Plant Working Group, 2009; Dev et al., 2014).

The primer sequences for *trnH-psbA* and *rbcL* were obtained from Tate and Simpson (2003), and Kress *et al.* (2009), respectively. The primers for *matK* were designed using Primer 3 software (Primer 3, version 4.1.0). The primer sequences can be viewed in Table S. 2. Polymerase chain reactions (PCR) were performed on the respective DNA extracts, followed by Sanger sequencing. The PCR master-mix for one reaction contained the following: 1.25 µl forward primer (10 µM), 1.25 µl reverse primer (10 µM), 8 µl ddH$_2$O, 12.5 µl KAPA2G fast HotStart Readymix, and 2 µl template DNA (100 ng).

Thermocycling reactions were performed following the standard protocol set out by KAPA Biosystems for the KAPA2G Fast HotStart ReadyMix kit with minor modifications (Roche Sequencing and Life Science, United States). The initial denaturation of the DNA was set at 95°C for 3 min followed by 25 cycles of denaturation (95°C for 15 s), annealing (60°C for 15 s), and extension (72°C for 15 s), with a final DNA extension step set at 72°C for 60 s. The reaction mixtures were then cooled down and kept at 4°C until further analysis.

The PCR products were cleaned using Sephadex G-50 (Sigma-Aldrich, South Africa) and Macherey-Nigel NucleoSpin® (MACHEREY-NAGEL GmbH & Co. KG, Germany) columns in preparation for cycle sequencing PCR. Each cycle sequencing reaction mixture was made up of 1 µl BigDye v3.1, 1.5 µl 5x Sequencing Buffer, 1.6 µl primer (forward and reverse reactions were done in separate sequencing PCR tubes), 2.9 µl dd H_2O, and 3 µl clean PCR amplified template DNA. This reaction mixture was placed in a thermocycler (Thermofisher, United States) set to the following thermal program; initial denaturation (95°C for 60 s) followed by 25 cycles of denaturation (95°C for 10 s), annealing (50°C for 5 s) and extension (60°C for 4 min), after which the reaction was cooled down and kept at 4°C until the samples could be sequenced at the sequencing facility of the University of Pretoria (Seqserve, Bioinformatics and Computational Biology Unit).

2.3.3. DNA barcoding: DNA barcode processing

The obtained sequences were processed using Geneious Prime bioinformatics software (version 2020.1.2; Biomatters Ltd., New Zealand). The forward and reverse sequences were aligned using the Geneious global pairwise alignment tool with free end gaps, followed by manual editing of ambiguous bases and trimming the poor-quality sequence ends, thus generating a consensus sequence for each individual tree's respective barcode.

The default alignment parameters included an open gap penalty of 10, gap extension penalty set to 3 and the refinement iterations equal to 10. The obtained sequence data have been deposited onto the National Centre for Biotechnology Information (NCBI) database (see Table S. 1. for the GenBank accession numbers). The sequence to species validation was done using the MegaBLAST functionality on the NCBI database.

After processing and validation, the barcodes were joined to generate a consensus sequence containing *matK* + *rbcL* + *trnH-psbA* for each tree. The consensus sequences were imported to SIMCA-P for qualitative data analyses using centre scaled principal component analysis (PCA) (World, 1987) scatter plots and hierarchical cluster analysis (HCA) dendrograms.

Species clustering based on evolutionary relatedness was also assessed by utilising the Bayesian modelling plugin v3.2.6 on Geneious Prime (Altekar et al., 2004; Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) as well as the maximum likelihood (ML) inference plugin PhyML v3.0 (Guindon et al., 2010; Guindon and Gascuel, 2003). The nucleotide substitution model for both phylogenetic analyses were determined using jModelTest2 software (Darriba et al., 2012); F81 (Felsenstein, 1981), with rate variation among sites returned as the best suited model for the sequence data provided, supported by the deltaBIC (Bayesian informative criterion) and deltaDT (Decision theory criterion) scores. Additional parameters for the Bayesian analysis were set as follows; the rate of substitution variation was set to gamma (8 gamma categories), with the number of iterations equal to one million at a subsampling frequency of 1000, 4 heated chains and a burn-in length of 25%. The default priors parameters for the unconstrained branch lengths were used. Determining the best preforming ML tree regarding optimised topology, branch length and nucleotide substitution rates, the following parameters were set; rapid bootstrapping of 1000 bootstraps were used to evaluate branch support and based on the aligned sequence data, the proportion of invariable sites, as well as the gamma distribution were estimated. The number of substitution rate categories was set to four. A consensus tree was generated with a support threshold of 80% and a burn-in frequency of 25%. The sequence data representing the outgroup in both phylogenetic analyses (*Linum usitatissimum* L.) were obtained from the NCBI database (NC_036356.1) and processed similarly to the other sequence data used in this study. *Linum usitatissimum* (common flax) was chosen as the outgroup because of its phylogenetic relatedness to the study group (USDA and NRCS, 2021) and because of a lack in sequence data of other closely related genera within the Erythroxylaceae family, on sequence database repositories.

3. Results and discussion

3.1. Comparison of ^{1}H-NMR and GC-MS chromatogram fingerprints to determine the chemical relatedness of taxa

The chemotaxonomic-based ^{1}H-NMR analysis focused primarily on the polar metabolic compound fingerprint, while the GC-MS analysis revealed the extracted polar volatile metabolites within the species. The stacked ^{1}H-NMR spectra and GC-MS chromatograms (Fig. 1) showcase the high degree of chemical relatedness between the taxa. When focusing on distinct regions, such as the aromatic and halogen chemical shifts (δ5–δ8 ppm) of the ^{1}H-NMR spectra (Fig. 1, A), species-specific groupings become more apparent. *Nectaropetalum capense* and *N. zuluense* show greater genus-based similarity in their polar metabolic fingerprints (NMR) as compared to the *Erythroxylum* species. *Erythroxylum delagoense* and *E. pictum* share a higher degree of chemical relatedness; with *E. emarginatum* being visually distinguishable from the rest. These results are supported by the stacked GC-MS chromatograms (Fig. 1, B), where species-level differentiating characteristics are more pronounced within *Erythroxylum*, as compared to those of the two *Nectaropetalum* species, indicating a need for further investigation.

3.2. ^{1}H-NMR and GC-MS MVDA

The results from the stacked chromatograms correlate well to the SIMCA-P generated ^{1}H-NMR and GC-MS PCA/ HCA dendrograms and OPLS-DA score plots (Fig. 2). The PCA/ HCA dendrograms (Fig. 2, A) show significant genus-based separation (A1, R^2(*cum*) – 0.73 & Q^2(*cum*) – 0.56; A2, R^2(*cum*) – 0.84 & Q^2(*cum*) – 0.48), while providing species-level resolution only within *Erythroxylum*. The supervised OPLS-DA score plots (Fig. 2, B) support these results; however, overlap between several replicates of *E. delagoense* and *E. pictum* is observed. This clustering was found to result from the reduction in dimensional viewing when this two-dimensional plot was generated, as separation in the z-axis (t [3]) was observed in the 3D model (results not shown).

The two *Nectaropetalum* species share a high degree of chemical similarity (Fig. 2 A1 & A2) and could only be partially separated with the supervised OPLS-DA score plots (Fig. 2, B). This overlap between the two species could result from the subtle chemical differences being suppressed by the high degree of chemical similarity present within these species, thus reducing the differentiating capability of the models. Both OPLS-DA plots were found to be significant based on permutation (100 permutations: 4 components) and cross validation (CV)-ANOVA results (*p*-value < 0.05).

Assessing each genus separately resulted in a higher degree of species-level resolution (Fig. 3). The *Erythroxylum* ^{1}H-NMR and GC-MS PCA/ HCA dendrograms (Fig. 3, A1; C1) revealed good species-level clustering, supported by a Hotelling's T2 of 0.05, as well as showing reliability (R^2(*cum*)) and predictability (Q^2(*cum*)) scores within the limits for biological samples (Eriksson et al., 2013). These findings correspond to the permutation and CV-ANOVA validated OPLS-DA results (Fig. 3, A2 & C2). Indicating that the largest contributing factor to the differences observed between *E. delagoense* and *E. pictum* are due to differences in functional group concentrations (^{1}H-NMR), or compound peak intensities (GC-MS), rather than a difference in compound presence or absence. This can be seen with the variation on the y-axis of the OPLS-DA score plots (Wiklund, 2008), relating to peak intensities (GC-MS) or functional group concentration (^{1}H-NMR). *E. emarginatum* was once again easily distinguishable in both supervised and unsupervised models, where the contributing factors were found to be compound, as well as functional group differences, resulting in x- and y-axis differentiation (Fig. 3, A2 & C2).

The unsupervised analysis of the ^{1}H-NMR and GC-MS data relating to the *Nectaropetalum* species (Fig. 3, B1 & D1) were unable to provide sufficient differentiation. The failure in differentiating between the two species is exacerbated by the poor level of clustering observed from the individual replicates of the two species. This indicates a lack in pronounced and consistent differentiating characteristics. The supervised analysis of the *Nectaropetalum* genus (Fig. 3, B2 & D2) does show slight differentiation; although, the ^{1}H-NMR OPLS-DA plot was found not to be significant, based on the CV-ANOVA results, permutation analysis and Q^2 (*cum*) score of 0.17.

Furthermore, the slight differentiation achieved with the supervised GC-MS score plot (Fig. 3, D2) should also be viewed with caution, as the variation between the x-variables (peak presence and intensity) is extremely limited (R^2X (*cum*) = 0.39), indicating a large degree of overlap between the two species which could be masking the differentiating characteristics. Both OPLS-DA plots of the *Nectaropetalum* genus showed poor correlation to their respective PCA models, reducing the significance of the groupings observed (Worley and Powers, 2016).

3.2.1. Biomarker identification

As a result of the ^{1}H-NMR and GC-MS MVDA findings, an in-depth analysis was conducted following the process described in Fig. 4 using the GC-MS data. Multiple biomarkers were screened and only further evaluated if the specific biomarker was present in all replicates of the corresponding species, as well as obtaining an MS similarity match factor greater than 85% in each replicate.

A literature search was conducted on each potential biomarker to determine their biological functions, and the possibility of being artifacts based on the use of methanol as extraction solvent (Casale, 1992; Sauerschnig et al., 2018). This revealed three biomarkers per species (Table 1), each playing key roles as precursors in various biosynthetic pathways. These biomarkers could be used in future plant collections to validate the identity of the species studied.

The effect of seasonal variation on the selected populations should also be investigated, as these biomarkers could show seasonal fluctuations (Ahmed et al., 2012; Alberts et al., 2018; Sampaio et al., 2016). In addition, the geographical sample range should be broadened as geographical-based environmental conditions has been shown to affect the chemical profile of plants (Cramer et al., 2011; Endara et al., 2018). This suggests that the use of biomarkers to authenticate plant samples should be combined with other plant identification methods, limiting the possibility of false positive plant identification.

Based on their biological functions, the listed biomarkers (Table 1) are considered good candidates, because the survival of the plants is more likely to be influenced by their absence, than their small-scale seasonal and/or geographical fluctuations in concentration. The use of species-specific biomarkers could assist in the accuracy of intra-genus differentiation and is strongly advised when used in authenticating these species.

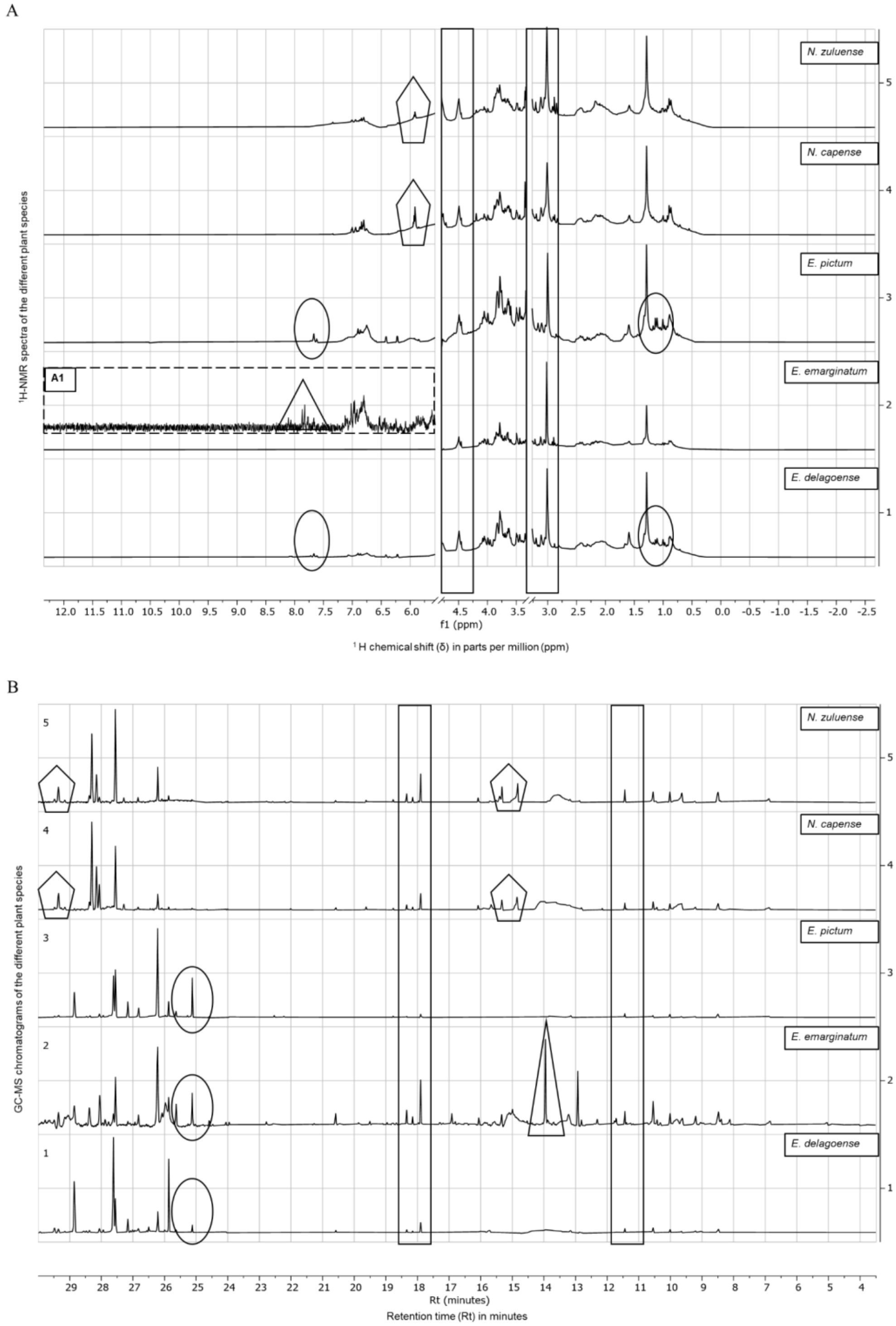


Fig. 1: Representative stacked spectra, and chromatograms. **A**, Proton – nuclear magnetic resonance (^{1}H-NMR) spectra of the five species assessed during this study. Residual CH$_3$OH solvent peaks have been removed. **A1**, Increased spectral intensity of E. emarginatum up-field from 5.5 ppm; **B**, Representative stacked gas chromatography – mass spectrometry (GC-MS) chromatograms of the five species assessed during the study. Selected similarity regions between all the species are shown in the boxes. Pentagons indicate similarity within only the *Nectaropetalum* genus, while similarities within *Erythroxylum* genus is indicated by ovals. The triangle shows peak differences that separate *E. emarginatum* from the other species. The stacked spectral images were generated using MestReNova software.

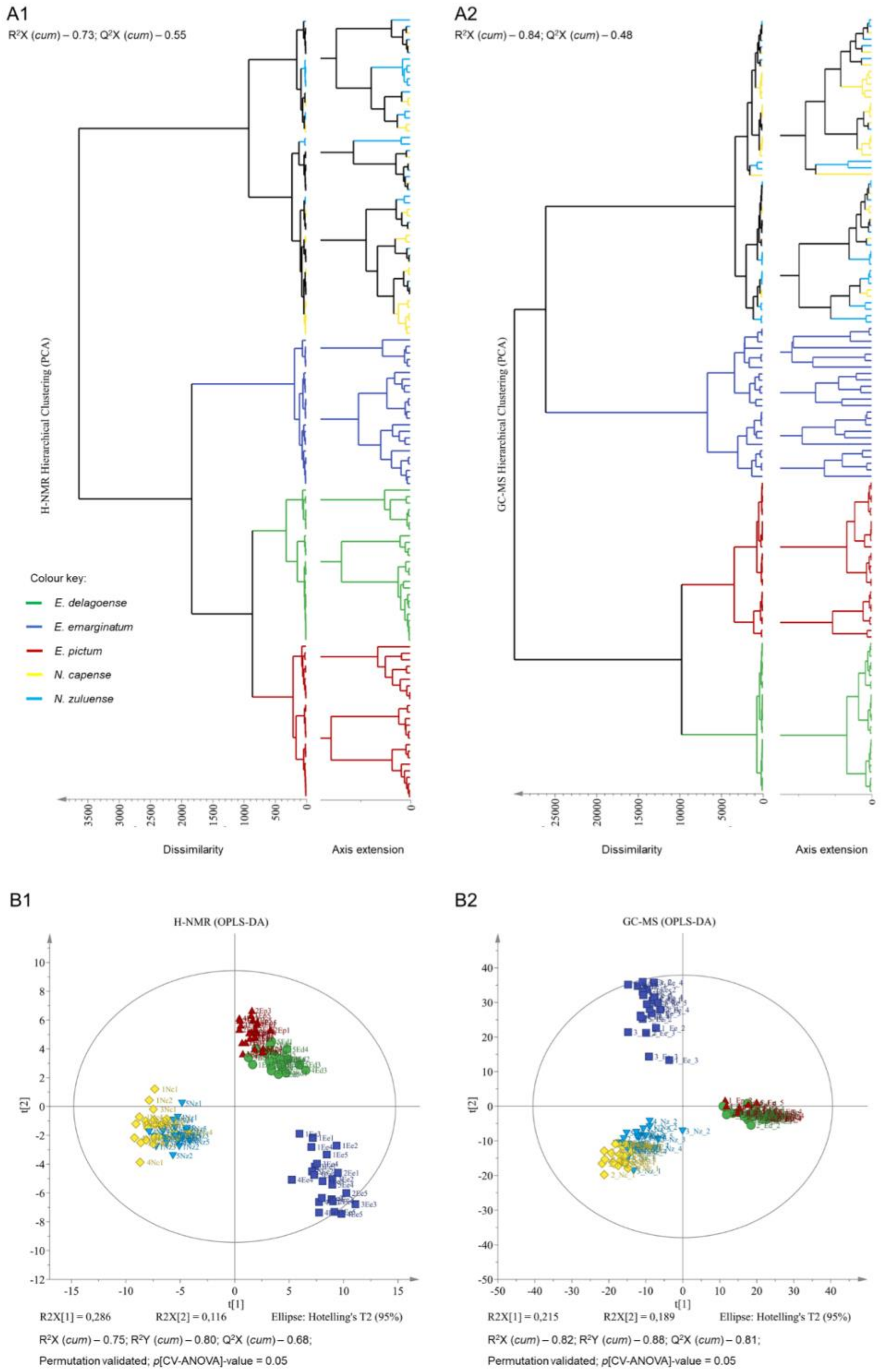


Fig. 2: The combined chemistries of the species assessed. Principal component (PC) based hierarchical cluster analysis (HCA) dendrograms of the proton - nuclear magnetic resonance (^{1}H-NMR) (**A1**) and gas chromatography - mass spectrometry (GC-MS) (**A2**) data obtained from all five species; **B1**, Orthogonal projections to latent structures discriminant analysis (OPLS-DA) of the ^{1}H-NMR and GC-MS (**B2**) data. The R^{2}X, Q^{2}X and R^{2}Y values represent the cumulative (cum) average of autofitted models. The OPLS-DA plots were generated with a Hotelling's T2 test of a 95% significance, validated by Permutation (100 permutations on the first four components) and subjected to cross validated (CV)-ANOVA significance testing. The plots and dendrograms were generated using the soft independent modelling by class analogy (SIMCA-P) software.

Table 1. Gas chromatography - mass spectrometry identified biomarkers of the south-east African Erythroxylaceae.

Plant taxon name	**Compound name**	**Molecular weight (amu)**	**Retention time (minutes)**	**Similarity index (% ± SD)**	**Previously identified *In Planta***
***Erythroxylum delagoense* Schinz**	Squalene	367	25.85–25.89	94±0.98	Lozano-Grande *et al.* (2018)
	1-Heptacosaol	396	27.61–27.63	97±0.09	Awaad *et al.* (2013); Kim *et al.* (2015)
	1,30-Tri-acontane-diol	454	28.84–28.86	91±0.49	Bhardwaj *et al.* (2018); Sampangi-Ramaiah *et al.* (2019)
***Erythroxylum emarginatum* Thonn.**	2-methoxy-3-prop-2-enylphenol	164	12.30–12.32	91±1.93	Wang *et al.* (2016); Ashraf *et al.* (2017)
	2-Carbo-methoxy-8-methyl-8-azabicyclo-[3.2.1]-oct-2-ene	181	12.91–12.95	90±0.63	Grynkiewicz & Gadzikowska, (2008); Kim *et al.* (2016)
	Ecgonine methyl ester	199	13.93–13.98	94±0.80	Grynkiewicz & Gadzikowska, (2008); Kim *et al.* (2016)

Table 1. (continued)

***Erythroxylum pictum* E.Mey. ex Harv. & Sond.**	1-Hentetracontanol	590	25.11–25.13	95±0.40	Tayade *et al.* (2013); Uritu *et al.* (2018)
	Tetratetracontane	589	26.19–26.24	95±0.98	Siddiqui *et al.* (2017)
	2-heptadecyloxirane	282	27.13–27.17	92±0.09	Dowd, (2012)
***Nectaropetalum capense* (Bolus) Stapf & Boodle**	Hexatriacontane	507	27.53–27.57	96±0.09	Barthlott *et al.* (2017); Soliman *et al.* (2019)
	Gamma-ergostenol	400	28.13–28.17	90±0.40	Quitain *et al.* (2001); Rossard *et al.* (2010)
	9,19-Cyclocholestan-3-ol	400	28.27–28.33	88±0.09	Achakzai *et al.* (2019)
***Nectaropetalum zuluense* (Schönland) Corbishley**	Beta-tocopherol	129	27.27–27.30	95±2.93	Fritsche *et al.* (2017); Mène-Saffrané, (2018)
	Gamma-ergostenol	416	28.13–28.16	90±1.02	Quitain *et al.* (2001); Rossard *et al.* (2010)
	9,19-Cyclocholestan-3-ol	400	28.29–28.31	88±0.49	Achakzai *et al.* (2019)

3.3. DNA barcoding

DNA barcoding was incorporated to address the challenge of chemical plasticity, as the genetic makeup of a particular species should be more stable under various environmental conditions. Especially when referring to the two conserved gene regions of *rbcL* and *matK* (CBOL Plant Working Group, 2009; Kress et al., 2005). The topology of the DNA based SIMCA-P generated PCA/ HCA dendrogram (Fig. 5, A1) corresponds to that of the phylogenetic trees, supported by good bootstrap (PhyML, Fig. 5, B1) and posterior probability (Bayesian, Fig. 5, B2) values. The dendrogram is not intended to be used as an evolutionary-based model. It is designed to assess variation in the dataset provided and does not consider other evolutionary related factors when generating the respective output results (Eriksson et al., 2013; Guindon et al., 2010; Ronquist and Huelsenbeck, 2003; World, 1987). However, it can be used as supportive data and a visual aid in determining which gene regions contribute the most to the variation between species, based purely on the sequence code. Results reveal that the PCA/ HCA dendrogram clearly differentiates between the two genera and indicates three distinct *Erythroxylum* species. The distinguishing factor being attributed to the intergenic spacer region *trnH-psbA* (Fig. 5, A2). Furthermore, the sequence variation observed within this locus demonstrates intra-species differentiation within *E. delagoense* and *E. pictum* (Fig. 5, A2). With greater evolutionary support being provided by the Bayesian analysis relating to *E. delagoense* (Fig. 5, B2). Intra-species variation can be expected when assessing *trnH-psbA*, as it is not a conserved gene region (compared to *rbcL* and *matK*). This allows for more gene variation to occur without deleterious effects to the plant (Aldrich et al., 1988; Bolson et al., 2015; Kress et al., 2005; Spooner, 2009).

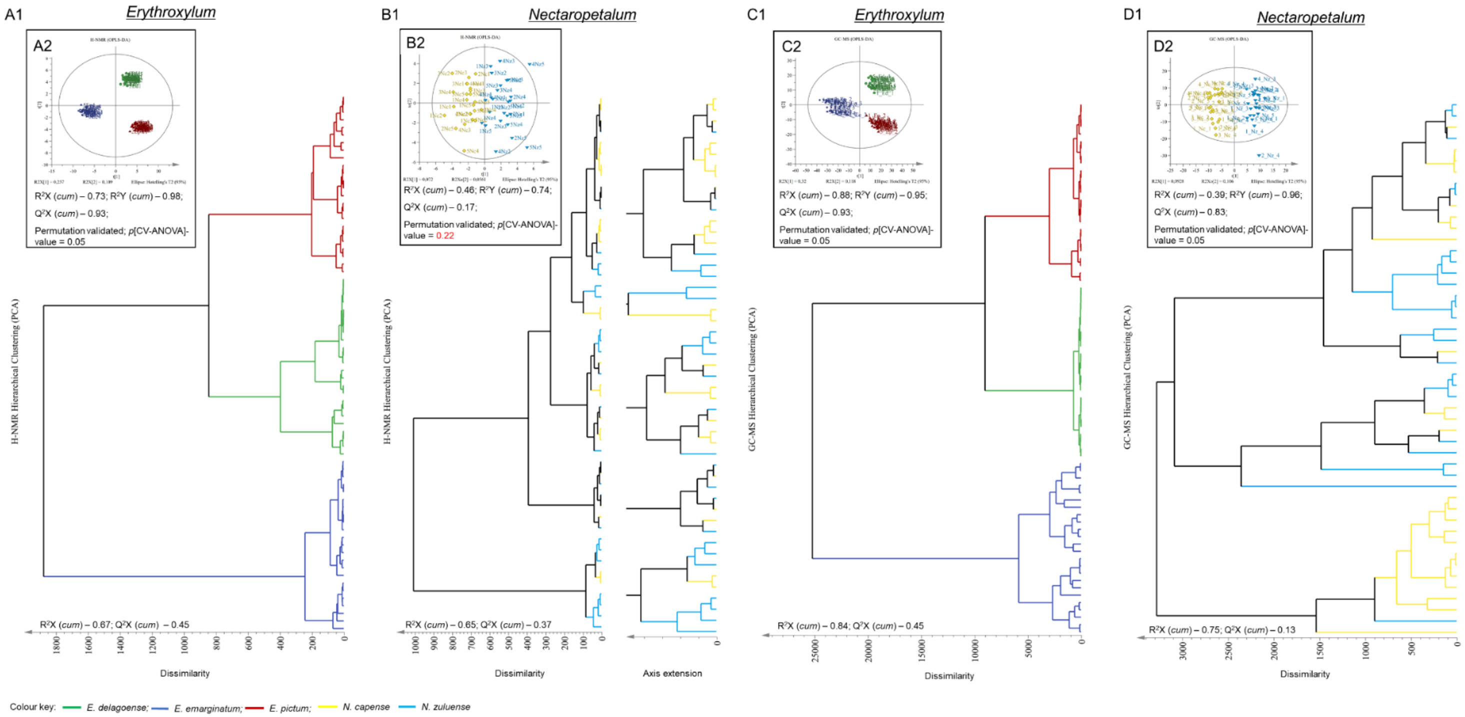
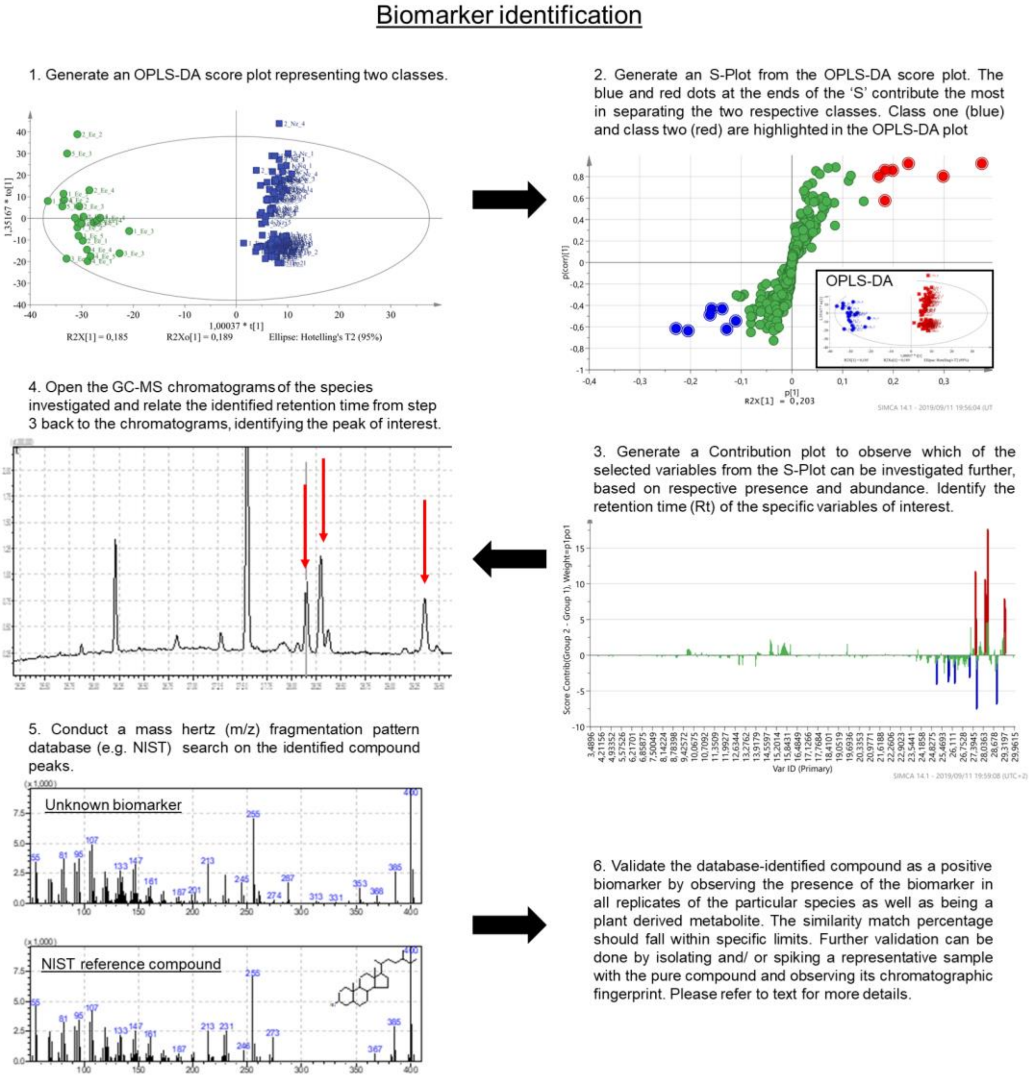


Fig. 3: Intra-generic variation of the species assessed. Genus-based principal component (PC) hierarchical cluster analysis (HCA) dendrograms (**A1; B1; C1; D1**) and orthogonal projections to latent structures discriminant analysis (OPLS-DA) score plots (**A2; B2; C2; D2**). The proton - nuclear magnetic resonance (^{1}H-NMR) analyses are represented in plots **A** and **B**, while plots **C** and **D** represent the gas chromatography - mass spectrometry analyses. The R^{2}X, Q^{2}X and R^{2}Y values represent the cumulative (cum) average of autofitted models. The OPLS-DA plots were generated with a Hotelling's T2 test of a 95% significance, validated by Permutation (100 permutations on the first four components) and subjected to cross validated (CV)-ANOVA significance testing. The plots were generated using the soft independent modelling by class analogy (SIMCA-P) software.



Biomarker identification

1. Generate an OPLS-DA score plot representing two classes.

2. Generate an S-Plot from the OPLS-DA score plot. The blue and red dots at the ends of the 'S' contribute the most in separating the two respective classes. Class one (blue) and class two (red) are highlighted in the OPLS-DA plot

3. Generate a Contribution plot to observe which of the selected variables from the S-Plot can be investigated further, based on respective presence and abundance. Identify the retention time (Rt) of the specific variables of interest.

4. Open the GC-MS chromatograms of the species investigated and relate the identified retention time from step 3 back to the chromatograms, identifying the peak of interest.

5. Conduct a mass hertz (m/z) fragmentation pattern database (e.g. NIST) search on the identified compound peaks.

6. Validate the database-identified compound as a positive biomarker by observing the presence of the biomarker in all replicates of the particular species as well as being a plant derived metabolite. The similarity match percentage should fall within specific limits. Further validation can be done by isolating and/ or spiking a representative sample with the pure compound and observing its chromatographic fingerprint. Please refer to text for more details.

Fig. 4: Flow diagram of the process used to identify potential biomarkers in each of the species. OPLS-DA – Orthogonal projections to latent structures discriminant analysis. GC-MS – Gas chromatography - mass spectrometry. NIST – National Institute for Standards and Technology.

Both phylogenetic models were able to form species-specific groupings (Fig. 5, B1 & B2), with greater species-level resolution seen from the Bayesian results. The PhyML analysis indicates that *E. pictum* is genetically indistinguishable from *E. delagoense* (Fig. 5, B1), and that between the three *Erythroxylum* species investigated, *E. emarginatum* is the most recently developed species. This lacks support from the Bayesian analysis (Fig

5, B2), showing *E. delagoense* as the most recently developed taxon although this species cluster seems unresolved with intra-species variation observed. The failed correlation between the two phylogenetic models could be attributed to the limited number genomic data assessed, as well as the different evolutionary-based model algorithms used in generating the respective output results (Guindon et al., 2010; Ronquist and Huelsenbeck, 2003). However, the collective findings from these phylogenetic models suggest that *E. delagoense* and *E. pictum* have recently diverged from a common ancestor, which would explain their high level of chemical and morphological relatedness.

These findings contrast with the extensive research done by Spooner (2009), who highlighted that *matK* and *trnH-psbA* are not suitable for differentiating certain species, as these barcodes fail to cluster well-defined taxa as a result of poor polymorphic representation. From the Erythroxylaceae investigated, *matK* and more notably *trnH-psbA* was able to provide species-level resolution between some species.

None of the loci assessed were able to differentiate between the two *Nectaropetalum* species. This, as well as the chemical relatedness suggests that the two *Nectaropetalum* species have also recently undergone evolutionary divergence. Further research into this matter is needed, as the inclusion of more genetic material such as the internal transcribed spacer region (ITS), may provide a greater degree of species resolution (Feliner and Rosselló, 2007; Kress, 2017; Li et al., 2011).

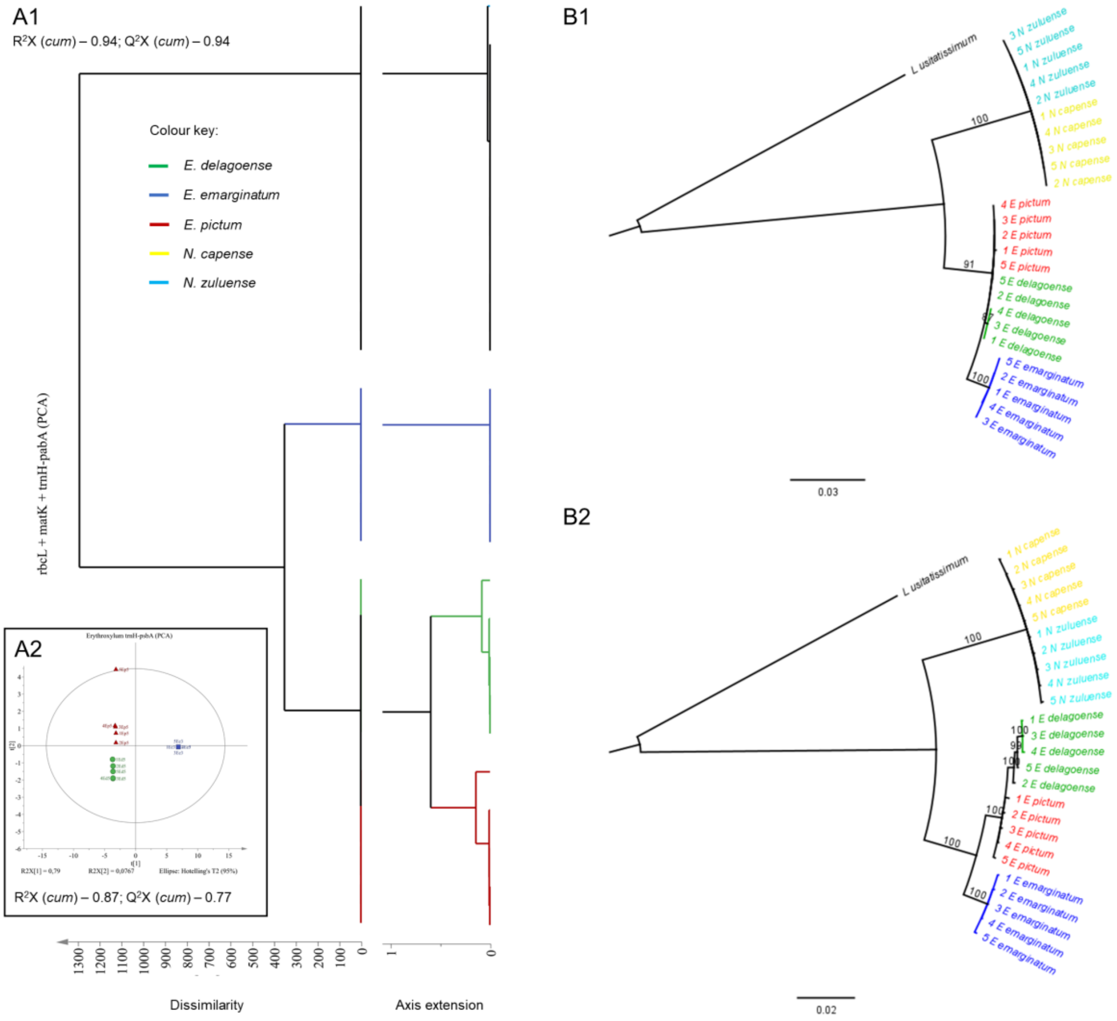


Fig. 5: DNA barcoding and small-scale phylogenetic analysis of the south-east African Erythroxylaceae. **A1,** Principal component (PC) based hierarchical cluster analysis (HCA) dendrogram of the DNA barcoding (*matK* + *rbcL* + *trnH-psbA*) data obtained from all five species belonging to the two genera *Erythroxylum* and *Nectaropetalum*. **A2,** Intrageneric related PC analysis (PCA) score plot of the *trnH-psbA* DNA barcode data from the *Erythroxylum* species. The R^2X and Q^2X values represent the cumulative (*cum*) average of autofitted models. The HCA dendrogram and PCA plot was generated using the soft independent modelling by class analogy (SIMCA-P) software. **B1,** Consensus maximum-likelihood phylogenetic analysis constructed using PhyML version 3.0, branch support indicated by the bootstrap values; **B2,** Consensus Bayesian phylogenetic analysis tree, branch support indicated by the posterior probability percentages. The phylogenetic trees were generated using the PhyML and MrBayes plugins in Geneious Prime software.

3.4. Comparing identification methodologies

The use of multiple plant identification methods is required in certain instances especially when the plants under study share a high level of morphological and/or molecular (DNA and chemical) similarity. However, there are many factors which need to be considered when deciding on which methodology to use, such as what material is available for analysis, the instrumental sensitivity and species differentiating capabilities, the purpose of the acquired data, the time available for analysis and what the financial implications of the different methods are. Recommendations of what instrumentation to use regarding the plant material available as well as the sensitivity of the different methodologies have been described (Alonso et al., 2015; Boggia et al., 2017; Boughton et al., 2016; Endara et al., 2018; Heyman and Meyer, 2012; Jorge et al., 2016; Kress, 2017; Mishra et al., 2016; Petruczynik, 2012). The financial aspect and time required for analysis are factors greatly overlooked and very rarely reported in literature. Thus, we aim to provide an overview of the financial implications and average time spent on sample preparation and analysis related to the methodologies described in this study (DNA barcoding, NMR and GC-MS). Furthermore, we compare the financial cost of these to other plant identification methodologies.

To generate an informed financial cost estimate of each method three cost categories were investigated: in-house, academic, and commercial rates for sample analysis. The academic and commercial rates were calculated with (full-service) and without (half-service) the cost of assisted data interpretation. All costs were based on the average of three prices per method from different analytical facilities, institutions and/ or universities, ranging from the years 2019 to 2021. The prices were all converted to the same currency i.e., South African Rand (ZAR) based on the currency exchange rates at the date of reporting (USD/ZAR = 14.73, EUR/ZAR = 17.37, and GBP/ZAR = 20.60), followed by calculating the average cost for 10 samples at in-house, academic, and commercial rates. The full cost calculations, products and reagents are provided as reference in the Supplementary material, as well as a summary of the experimental workflow and average time required for each of the three methods on which this study focused.

It should be noted that the cost calculations for the different methodologies are merely for comparative and research purposes. The financial cost calculations presented in this study were generated from available price lists at the time of reporting and can vary based on factors such as the professional relationship between the research group and institution, university, or analytical facility.

The average cost for commercial analysis far outweighs both academic and in-house rates, with in-house rates being the lowest. In-house GC-MS analysis was found to be the most cost effective at ZAR 2683.80 per 10 samples, followed by DNA barcoding (ZAR 4860.70) and ^{1}H-NMR (ZAR 5124.60). The academic rates (full-service) for DNA barcoding, ^{1}H-NMR and GC-MS were 184%, 208%, and 454% more expensive, respectively. The full-service commercial rates for DNA barcoding, ^{1}H-NMR, and GC-MS were 256%, 413%, and 662% more expensive, respectively.

When comparing the full-service commercial and academic rates, the commercial rates for ^{1}H-NMR were found to be almost 200% more expensive, followed by GC-MS 146% and DNA barcoding 139%. Focusing purely on the academic rates, DNA barcoding of three genetic markers was the least expensive at ZAR 8938.80 per 10 samples, with ^{1}H-

NMR and GC-MS being 119% and 136% more expensive. This highlights the drastic cost increase from in-house to academic to commercial rates.

Furthermore, when comparing the academic cost of these methods to other plant differentiating methodologies such as high performance liquid chromatography (HPLC) (Viapiana et al., 2016), liquid chromatography-mass spectrometry (LC-MS) (Kharyuk et al., 2018), matrix-assisted laser desorption/ionization-mass spectrometry (MALDI-MS) (Ahmad et al., 2012), ultraviolet-visible spectroscopy (UV-Vis) (Philippidis et al., 2021) and electrospray ionisation-mass spectrometry (ESI-MS) (Goodacre et al., 2003), then it becomes apparent that LC-MS is the most expensive analysis and UV-Vis the least expensive. The average academic rates for analysing 10 samples at 30 min per sample using UV-Vis equates to ZAR 3880.60, with HPLC being 134% more expensive, followed by MALDI-MS (186%), ESI-MS (215%), DNA barcoding (230%), ^{1}H-NMR (274%), GC-MS (314%), and LC-MS showing an increase of 334% in price.

Data interpretation and method development play major roles in determining the cost of analysis, especially related to chemotaxonomic and other chemometric approaches. When these factors are excluded from the cost calculations there is an average decrease of 137% compared to the full-service cost for academics. The half-service academic price for GC-MS analysis proves to be the costliest with a 298% price increase compared to the least expensive analysis (UV-Vis ZAR 3052.00). This highlights the financial implication related to assisted data interpretation, especially for MS related analyses.

One of the biggest drawbacks to conventional DNA barcoding using Sanger sequencing methodologies is the financial implication and time required for sample preparation and analysis. It can take an average of 199 hours to prepare and process three barcoding genes from 25 plant samples, where it would take approximately 110 hours to prepare and analyse the same number of samples with ^{1}H-NMR and GC-MS combined. However, the development of plant-based next-generation sequencing (NGS) technologies as well as novel DNA barcoding approaches could alleviate these drawbacks in the near future (Gostel et al., 2020; Stein et al., 2014).

Based on the financial cost calculations, the data reported, and the literature reviewed during this study, the recommended methodologies to use for plant identification and differentiation is dependent on the instrumentation available, the professional relationship between the research group and the analysis facility, the global purpose of the generated data for other downstream applications, the availability of reference databases as well as the availability of instrument specific analysis methods. Should the interest be based purely on identifying and differentiating between morphologically similar plant species, and the financial implication of the different methodologies at academic and commercial rates, not considering the aforementioned factors then UV-Vis, ^{1}H-NMR and DNA barcoding would be recommended. UV-Vis provides rapid cost-effective results as well as species-specific UV absorption patterns (Boggia et al., 2017). ^{1}H-NMR can provide a global metabolic fingerprint of the more abundant plant metabolites, which can be used in downstream applications such as compound functional group analysis (Sumner et al., 2003), whereas DNA barcoding contributes to the genetic-based differentiating aspects and provides data that can be used in downstream phylogenetic reconstructions (Kang et al., 2017). If financial cost is not a consideration, then the use of all three methods described in this study would be recommended, therefore replacing UV-Vis with GC-MS as it can provide a higher degree of differentiating sensitivity coupled to well established MS compound library databases.

3.5. Discussion

The difficulty in distinguishing between closely related species such as the Erythroxylaceae species investigated in this study, is a common problem in biology. Mishra et al. (2016) found during their comprehensive review relating to DNA barcoding and the authentication challenges of plants at herbal markets, that the integrated use of DNA barcoding with "-omics" (proteomics, metabolomics and transcriptomics) systems, should be used for the identification and differentiation of plant species. Furthermore, Endara et al. (2018) described how the use of chemotaxonomy was superior in differentiating between closely related species, compared to the morphological and DNA barcoding techniques assessed during their study. The financial cost comparison showed that the techniques described in this study (^{1}H-NMR, GC-MS and DNA barcoding) are some of the more expensive techniques at academic and commercial rates. However, the in-house rates and rapid results indicate that the use of these methods for plant identification and differentiation is financially feasible, especially when considering the data generated by these techniques and the potential downstream applications.

Our results show that the combined use of chemotaxonomy with DNA barcoding was able to differentiate between the species of Erythroxylaceae investigated. However, the chemical plasticity observed in this study needs to be kept in mind during future sample collection and authentication, as seasonal variation and geographical location of the plants may alter the chemical-based species groupings observed. This, and the factors mentioned from previous studies highlight the need for a more integrated plant identification technique. The combined differentiating capabilities of morphology with chemotaxonomy and DNA barcoding can prove to be useful in identifying and separating other closely related and/ or morphologically similar taxa. This can especially be the case when sample populations are small, the genomic information limited, or where the effect of different environmental stimuli can alter the chemical makeup of the sample and thus affect the identification potential of these techniques.

4. Concluding remarks

Accurate and reliable identification of closely related and morphologically similar medicinal plants is crucial, as misidentification could have dire consequences. The *Artemisia* genus can be seen as such an example, where *A. annua* L. and *A. afra* Jacq. Ex Willd. are morphologically nearly indistinguishable. However, chemically *A. annua* contains the active compound artemisinin (used in malaria treatment), whereas *A. afra* does not (Van der Kooy et al., 2008). This can result in a fatal outcome if the incorrect plant is used for the treatment of malaria and related symptoms.

We have shown that by incorporating both facets of DNA barcoding and sensitive chemotaxonomic methods, the misidentification of morphologically similar species, such as the species investigated in this study, could be resolved. The financial aspect related to the chosen methodologies for this study is a factor which needs to be considered. However, based on the in-house cost, results from this study and the many possible downstream applications of the generated data (not limited to only plant identification and classification), we trust that these methodologies have shown their separate and combined feasibility in species differentiation.

To our knowledge, the use of NMR- and GC-MS- based metabolomics in combination with DNA barcoding has not previously been described for the identification and differentiation of plant species. The findings from this study compare well with previous reports related to the integrated use of identification methodologies (Bentley et al., 2019;

Endara et al., 2018; Kress et al., 2005; Mishra et al., 2016), while furthermore providing financial and time-based comparative aspects.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Acknowledgements

We would like to thank the members of the CREW (Custodians of Rare and Endangered Wildflowers) program from the Umtamvuna Nature Reserve and Mr P. Alberts for assisting with the plant collections. Prof. A.E. van Wyk, Mrs E.P. van Wyk and Ms M. Nel of the H.G.W.J. Schweickerdt Herbarium of the University of Pretoria are thanked for assisting with the identification of the plants and Mrs N.L. Meyer for assisting with the overall grammar and language editing of the manuscript. The authors would further like to extend their gratitude to the assigned Editor (Kevin Balkwill, PhD) and anonymous Reviewers for taking the time to carefully assess the work presented in this study, providing valued comments and corrections. We appreciate the Ph.D. funding support (Innovation and Scares Skills) provided to P.S.F Alberts by The National Research Foundation (Grant UID: 110698). The research data presented in this manuscript did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

The following additional information relating to this study can be found in the Supplementary Material (see online published version):

Fig. S. 1. The mass fragmentation patterns of the identified biomarkers and library references.

Fig. S. 2. Flow diagram describing the main experimental and analysis steps used, and the average time required (in an ideal scenario) to complete each section.

Table S. 1. The collection sites, PRU numbers and GenBank accession numbers of the plant species assessed.

Table S. 2. DNA primer sequences used in the study.

Table S. 3. Summary of the financial cost comparison between different analytical techniques used in plant identification.

References

Abascal, K., Ganora, L., Yarnell, E., 2005. The effect of freeze-drying and its implications for botanical medicine: A review. Phytotherapy Research 19, 655–660. https://doi.org/10.1002/ptr.1651

Achakzai, J.K., Anwar Panezai, M., Kakar, A.M., Akhtar, B., Akbar, A., Kakar, S., Khan, J., Khan, N.Y., Khan, G.M., Baloch, N., Jan Khoso, M.H., Panezai, M., 2019. In vitro antileishmanial activity and GC-MS analysis of whole plant hexane fraction of *Achillea wilhelmsii* (WHFAW). Journal of Chemistry 2019. https://doi.org/10.1155/2019/5734257

Afzan, A., Bréant, L., Bellstedt, D.U., Grant, J.R., Queiroz, E.F., Wolfender, J.L., Kissling, J., 2019. Can biochemical phenotype, obtained from herbarium samples, help taxonomic decisions? – A case study using Gentianaceae. Taxon 68, 771–782. https://doi.org/10.1002/tax.12120

Ahmad, F., Babalola, O.O., Tak, H.I., 2012. Potential of MALDI-TOF mass spectrometry as a rapid detection technique in plant pathology: Identification of plant-associated microorganisms. Analytical and Bioanalytical Chemistry 404, 1247–1255. https://doi.org/10.1007/s00216-012-6091-7

Ahmad, M., Khan, M.A., Rashid, U., Zafar, M., Arshad, M., 2009. Quality assurance of herbal drug valerian by chemotaxonomic markers. Journal of Biotechnology 8, 1148–1154.

Ahmad, M., Khan, M.A., Zafar, M., Arshad, M., Sultana, S., Abbasi, B.H., 2010. Use of chemotaxonomic markers for misidentified medicinal plants used in traditional medicines. Journal of Medicinal Plants Research 4, 1244–1252.

https://doi.org/10.5897/JMPR10.027

Ahmed, D., Baig, H., Zara, S., 2012. Seasonal variation of phenolics, flavonoids, antioxidant and lipid peroxidation inhibitory activity of methanolic extract of *Melilotus indicus* and its sub-fractions in different solvents. International Journal of Phytomedicine 4, 326–332.

Alberts, P.S.F., Daneel, M., Marais, A.A.S., Baranenko, D.A., Meyer, J.J.M., 2018. Seasonal analysis of the tropane alkaloid ecgonine methyl ester and the occurrence of other highly-valued tropanes in the South African *Erythroxylum* trees. Acta Physiologiae Plantarum 40. https://doi.org/10.1007/s11738-017-2599-y

Aldrich, J., Cherney, B.W., Merlin, E., 1988. The role of insertions/deletions in the evolution of the intergenic region between *psbA* and *trnH* in the chloroplast genome. Current Genetics 14, 137–146. https://doi.org/10.1007/BF00569337

Alonso, A., Marsal, S., Julià, A., 2015. Analytical methods in untargeted metabolomics: State of the art in 2015. Frontiers in Bioengineering and Biotechnology 3, 1–20. https://doi.org/10.3389/fbioe.2015.00023

Altekar, G., Dwarkadas, S., Huelsenbeck, J.P., Ronquist, F., 2004. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics 20, 407–415. https://doi.org/10.1093/bioinformatics/btg427

Ashraf, S.A., Al-Shammari, E., Hussain, T., Tajuddin, S., Panda, B.P., 2017. In-vitro antimicrobial activity and identification of bioactive components using GC–MS of commercially available essential oils in Saudi Arabia. Journal of Food Science and Technology 54, 3948–3958. https://doi.org/10.1007/s13197-017-2859-2

Awaad, A.S., El-Meligy, R.M., Al-Jaber, N.A., Al-Muteeri, H.S., Zain, M.E., Alqasoumi, S.I., Alafeefy, A.M., Donia, A.E.R.M., 2013. Anti-ulcerative colitis activity of compounds from *Euphorbia granuleta* Forssk. Phytotherapy Research 27, 1729–1734. https://doi.org/10.1002/ptr.4985

Barthlott, W., Mail, M., Bhushan, B., Koch, K., 2017. Plant surfaces: Structures and functions for biomimetic innovations. Nano-Micro Letters 9, 1–40. https://doi.org/10.1007/s40820-016-0125-1

Bentley, J., Moore, J.P., Farrant, J.M., 2019. Metabolomics as a complement to phylogenetics for assessing intraspecific boundaries in the desiccation-tolerant medicinal shrub *Myrothamnus flabellifolia* (Myrothamnaceae). Phytochemistry 159, 127–136. https://doi.org/10.1016/j.phytochem.2018.12.016

Bhardwaj, P., Bhardwaj, G., Raghuvanshi, R., Thakur, M.S., Kumar, R., Chaurasia, O.P., 2018. Rhodiola: An overview of phytochemistry and pharmacological applications, New Age Herbals: Resource, Quality and Pharmacognosy. https://doi.org/10.1007/978-981-10-8291-7_5

Bhatta, S., Janezic, T.S., Ratti, C., 2020. Freeze-drying of plant-based foods. Foods 9, 1–22. https://doi.org/10.3390/foods9010087

Boggia, R., Turrini, F., Anselmo, M., Zunin, P., Donno, D., Beccaro, G.L., 2017. Feasibility of UV–VIS–Fluorescence spectroscopy combined with pattern recognition techniques to authenticate a new category of plant food supplements. Journal of Food Science and Technology 54, 2422–2432. https://doi.org/10.1007/s13197-017-2684-7

Bolson, M., De Camargo Smidt, E., Brotto, M.L., Silva-Pereira, V., 2015. *ITS* and t*rnH-psbA* as efficient DNA barcodes to identify threatened commercial woody angiosperms from southern Brazilian Atlantic rainforests. PLoS ONE 10, 1–18. https://doi.org/10.1371/journal.pone.0143049

Boon, R., 2010. Pooley’s Trees of Eastern South Africa, Second. ed. Flora and Fauna Publications Trust.

Boughton, B.A., Thinagaran, D., Sarabia, D., Bacic, A., Roessner, U., 2016. Mass spectrometry imaging for plant biology: a review. Phytochemistry Reviews 15, 445–488. https://doi.org/10.1007/s11101-015-9440-2

Casale, J.F., 1992. Methyl Esters of Ecgonine: Injection-Port Produced Artifacts from Cocaine Base (Crack) Exhibits. Journal of Forensic Sciences 37, 13317J. https://doi.org/10.1520/JFS13317J

CBOL Plant Working Group, 2009. A DNA barcode for land plants. Proceedings of the National Academy of Sciences of the United States of America 106, 12794–7. https://doi.org/10.1073/pnas.0905845106

Chase, M.W., Hills, H.H., 1991. Silica Gel: An ideal material for field preservation of leaf samples for DNA studies. Taxon 40, 215–220.

Coates-Palgrave, K., 2002. Keith Coates-Palgrave Trees of southern Africa, 3rd ed. Struik Publishers, Cape Town.

Corrigan, B.M., Van Wyk, B.E., Geldenhuys, C.J., Jardine, J.M., 2011. Ethnobotanical plant uses in the KwaNibela Peninsula, St Lucia, South Africa. South African Journal of Botany 77, 346–359. https://doi.org/10.1016/j.sajb.2010.09.017

Cramer, G.R., Urano, K., Delrot, S., Pezzotti, M., Shinozaki, K., 2011. Effects of abiotic stress on plants: a systems biology perspective. BMC Plant Biology 11, 163. https://doi.org/10.1186/1471-2229-11-163

Cullings, K.W., 1992. Design and testing of a plant-specific PCR primer for ecological and evolutionary studies. Molecular Ecology 1, 233–240. https://doi.org/10.1111/j.1365-294X.1992.tb00182.x

Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2012. JModelTest 2: More models, new heuristics and parallel computing. Nature Methods 9, 772. https://doi.org/10.1038/nmeth.2109

de Lima, R.A.F., de Oliveira, A.A., Colletta, G.D., Flores, T.B., Coelho, R.L.G., Dias, P., Frey, G.P., Iribar, A., Rodrigues, R.R., Souza, V.C., Chave, J., 2018. Can plant DNA barcoding be implemented in species-rich tropical regions? A perspective from São Paulo State, Brazil. Genetics and Molecular Biology 41, 661–670. https://doi.org/10.1590/1678-4685-gmb-2017-0282

de Vere, N., Rich, T.C.G., Ford, C.R., Trinder, S.A., Long, C., Moore, C.W., Satterthwaite, D., Davies, H., Allainguillaume, J., Ronca, S., Tatarinova, T., Garbett, H., Walker, K., Wilkinson, M.J., 2012. DNA barcoding the native flowering plants and conifers of wales. PLoS ONE 7, 1–12. https://doi.org/10.1371/journal.pone.0037945

de Wet, H., 2011. Antibacterial activity of the five South African Erythroxylaceae species. African Journal of Biotechnology 10, 11511–11514.

Dev, S.A., Muralidharan, E.M., Sujanapal, P., Balasundaran, M., 2014. Identification of market adulterants in East Indian sandalwood using DNA barcoding. Annals of Forest Science 71, 517–522. https://doi.org/10.1007/s13595-013-0354-0

Dowd, M.K., 2012. Identification of the unsaturated heptadecyl fatty acids in the seed oils of *Thespesia populnea* and *Gossypium hirsutum*. JAOCS, Journal of the

American Oil Chemists’ Society 89, 1599–1609. https://doi.org/10.1007/s11746-012-2071-5

Doyle, J.J., Dickson, E.E.E., 1987. Preservation of plant smaples for DNA restriction endonuclease analysis. Taxon 36, 715–722.

Doyle, J.J., Doyle, J.L., 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bulletin 19, 11–15. https://doi.org/10.2307/4119796

Endara, M.J., Coley, P.D., Wiggins, N.L., Forrister, D.L., Younkin, G.C., Nicholls, J.A., Pennington, R.T., Dexter, K.G., Kidner, C.A., Stone, G.N., Kursar, T.A., 2018. Chemocoding as an identification tool where morphological- and DNA-based methods fall short: *Inga* as a case study. New Phytologist 218, 847–858. https://doi.org/10.1111/nph.15020

Eriksson, L., Byrne, T., Johansson, E., Trygg, J., Vikström, C., 2013. Multi- and Megavariate Data Analysis, in: Technometrics. pp. 323–355. https://doi.org/10.1198/tech.2003.s162

Feliner, G.N., Rosselló, J.A., 2007. Better the devil you know? Guidelines for insightful utilization of nrDNA *ITS* in species-level evolutionary studies in plants. Molecular Phylogenetics and Evolution 44, 911–919. https://doi.org/10.1016/j.ympev.2007.01.013

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution 17, 368–376. https://doi.org/10.1007/BF01734359

Fritsche, S., Wang, X., Jung, C., 2017. Recent advances in our understanding of tocopherol biosynthesis in plants: An overview of key genes, functions, and breeding of vitamin E improved crops. Antioxidants 6. https://doi.org/10.3390/antiox6040099

Fusco, G., Minelli, A., 2010. Phenotypic plasticity in development and evolution: Facts and concepts. Philosophical Transactions of the Royal Society B: Biological Sciences 365, 547–556. https://doi.org/10.1098/rstb.2009.0267

Goodacre, R., York, E. V., Heald, J.K., Scott, I.M., 2003. Chemometric discrimination of unfractionated plant extracts analyzed by electrospray mass spectrometry. Phytochemistry 62, 859–863. https://doi.org/10.1016/S0031-9422(02)00718-5

Gostel, M.R., Zúñiga, J.D., Kress, W.J., Funk, V.A., Puente-Lelievre, C., 2020. Microfluidic Enrichment Barcoding (MEBarcoding): a new method for high throughput plant DNA barcoding. Scientific Reports 10, 1–13. https://doi.org/10.1038/s41598-020-64919-z

Govaerts, R., 2001. How many species of seed plants are there? Taxon. https://doi.org/10.2307/1224723

Gratani, L., 2014. Plant phenotypic plasticity in response to environmental factors. Advances in Botany 2014, 1–17. https://doi.org/10.1155/2014/208747

Grynkiewicz, G., Gadzikowska, M., 2008. Tropane alkaloids as medicinally useful natural products and their synthetic derivatives as new drugs. Pharmacological Reports 60, 439–463.

Guindon, S., Dufayard, J., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New Algorithms and Methods to Estimate Maximim-Likelihood Phylogenies Assessing the Performance of PhyML 3.0. Systematic Biology 59, 307–321.

Guindon, S., Gascuel, O., 2003. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Systematic Biology 52, 696–704. https://doi.org/10.1080/10635150390235520

Hasan, Q.H., 2012. Tropane (MOTM June 2012) [WWW Document]. The Molecule of the Month URL http://www.chm.bris.ac.uk/motm/motm.htm#june2012 (accessed 4.23.30).

Heyman, H.M., Meyer, J.J.M., 2012. NMR-based metabolomics as a quality control tool for herbal products. South African Journal of Botany 82, 21–32. https://doi.org/10.1016/j.sajb.2012.04.001

Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics Applications Note 17, 754–755.

Johnson, E.L., Schmidt, W.F., Norman, H.A., 1998. Flavonoids as markers for *Erythroxylum* Taxa. Biochemical Systematics and Ecology 26, 743–759. https://doi.org/10.1016/S0305-1978(98)00042-8

Jorge, T.F., Rodriques, J.A., Caldana, C., Schmidt, R., van Dongen, J.T., Tomas-Oates, J., Antonio, C., 2016. Mass spectrometry-based plant metabolomics: Metabolite responces to abiotic stress. Mass Spectrometry Reviews 35, 620–649. https://doi.org/10.1002/mas.21449

Kang, Y., Deng, Z., Zang, R., Long, W., 2017. DNA barcoding analysis and phylogenetic relationships of tree species in tropical cloud forests. Scientific Reports 7, 1–9. https://doi.org/10.1038/s41598-017-13057-0

Kharyuk, P., Nazarenko, D., Oseledets, I., Rodin, I., Shpigun, O., Tsitsilin, A., Lavrentyev, M., 2018. Employing fingerprinting of medicinal plants by means of

LC-MS and machine learning for species identification task. Scientific Reports 8, 1–12. https://doi.org/10.1038/s41598-018-35399-z

Kim, N., Estrada, O., Chavez, B., Stewart, C., D'Auria, J.C., 2016. Tropane and granatane alkaloid biosynthesis: A systematic analysis. Molecules 21, 1–25. https://doi.org/10.3390/molecules21111510

Kim, T.J., Lee, K.B., Baek, S.A., Choi, J., Ha, S.H., Lim, S.H., Park, S.Y., Yeo, Y., Park, S.U., Kim, J.K., 2015. Determination of lipophilic metabolites for species discrimination and quality assessment of nine leafy vegetables. Journal of the Korean Society for Applied Biological Chemistry 58, 909–918. https://doi.org/10.1007/s13765-015-0119-6

Kress, W.J., 2017. Plant DNA barcodes: Applications today and in the future. Journal of Systematics and Evolution 55, 291–307. https://doi.org/10.1111/jse.12254

Kress, W.J., Erickson, D.L., Jones, F.A., Swenson, N.G., Perez, R., Sanjur, O., Bermingham, E., 2009. Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. Proceedings of the National Academy of Sciences of the United States of America 106, 18621–18626. https://doi.org/10.1073/pnas.0909820106

Kress, W.J., Wurdack, K.J., Zimmer, E.A., Weigt, L.A., Janzen, D.H., 2005. Use of DNA barcodes to identify flowering plants. Proceedings of the National Academy of Sciences of the United States of America 102, 8369–8374.

Li, D.Z., Gao, L.M., Li, H.T., Wang, H., Ge, X.J., Liu, J.Q., Chen, Z.D., Zhou, S.L., Chen, S.L., Yang, J.B., Fu, C.X., Zeng, C.X., Yan, H.F., Zhu, Y.J., Sun, Y.S., Chen, S.Y., Zhao, L., Wang, K., Yang, T., Duan, G.W., 2011. Comparative analysis of a large dataset indicates that internal transcribed spacer (*ITS*) should be incorporated into the core barcode for seed plants. Proceedings of the National Academy of Sciences of the United States of America 108, 19641–19646. https://doi.org/10.1073/pnas.1104551108

Lozano-Grande, M.A., Gorinstein, S., Espitia-Rangel, E., Dávila-Ortiz, G., Martínez-Ayala, A.L., 2018. Plant sources, extraction methods, and uses of squalene. International Journal of Agronomy 2018. https://doi.org/10.1155/2018/1829160

Mantuano, D.G., Barros, C.F., Scarano, F.R., 2006. Leaf anatomy variation within and between three “restinga” populations of *Erythroxylum ovalifolium* Peyr. (Erythroxylaceae) in Southeast Brazil. Revista Brasileira de Botanica 29, 209–215. https://doi.org/10.1590/S0100-84042006000200002

Mène-Saffrané, L., 2018. Vitamin E biosynthesis and its regulation in plants. Antioxidants 7, 1–17. https://doi.org/10.3390/antiox7010002

Mishra, P., Kumar, A., Nagireddy, A., Mani, D.N., Shukla, A.K., Tiwari, R., Sundaresan, V., 2016. DNA barcoding: An efficient tool to overcome authentication challenges in the herbal market. Plant Biotechnology Journal 14, 8–21. https://doi.org/10.1111/pbi.12419

Mora, C., Tittensor, D.P., Adl, S., Simpson, A.G.B., Worm, B., 2011. How many species are there on earth and in the ocean? PLoS Biology 9, e1001127. https://doi.org/10.1371/journal.pbio.1001127

Nishiyama, Y.Y., Moriyasu, M., Ichimaru, M., Sonoda, M.M., Iwasa, K., Kato, A., Juma, F.D., Mathenge, S.G., Mutiso, P.B.C., Chalo Mutiso, P.B., 2007. Tropane alkaloids from *Erythroxylum emarginatum*. Journal of Natural Medicines 61, 56–

58. https://doi.org/10.1007/s11418-006-0018-6

Petruczynik, A., 2012. Analysis of alkaloids from different chemical groups by different liquid chromatography methods. Central European Journal of Chemistry 10, 802–835. https://doi.org/10.2478/s11532-012-0037-y

Philippidis, A., Poulakis, E., Kontzedaki, R., Orfanakis, E., Symianaki, A., Zoumi, A., Velegrakis, M., 2021. Application of ultraviolet-visible absorption spectroscopy with machine learning techniques for the classification of cretan wines. Foods 10. https://doi.org/10.3390/foods10010009

Quitain, A.T., Oro, K., Katoh, S., Moriyoshi, T., 2001. Ethanol-modified supercritical carbon dioxide extraction of high-value oil from Okara. Japan.

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574. https://doi.org/10.1093/bioinformatics/btg180

Rossard, S., Roblin, G., Atanassova, R., 2010. Ergosterol triggers characteristic elicitation steps in *Beta vulgaris* leaf tissues. Journal of Experimental Botany 61, 1807–1816. https://doi.org/10.1093/jxb/erq047

Sampaio, B.L., Edrada-Ebel, R., Da Costa, F.B., 2016. Effect of the environment on the secondary metabolic profile of *Tithonia diversifolia*: A model for environmental metabolomics of plants. Scientific Reports 6, 1–11. https://doi.org/10.1038/srep29265

Sampangi-Ramaiah, M.H., Ravishankar, K.V., Shivashankar, K.S., Roy, T.K., Rekha, A., Hunashikatti, L.R., 2019. Developmental changes in the composition of leaf cuticular wax of banana influenced by wax biosynthesis gene expression: a case study in *Musa acuminata* and *Musa balbisiana*. Acta Physiologiae Plantarum 41. https://doi.org/10.1007/s11738-019-2934-6

Sauerschnig, C., Doppler, M., Bueschl, C., Schuhmacher, R., 2018. Methanol generates numerous artifacts during sample extraction and storage of extracts in metabolomics research. Metabolites 8, 1–19. https://doi.org/10.3390/metabo8010001

Scotland, R.W., Wortley, A.H., 2003. How many species of seed plants are there? Taxon 52, 101–104. https://doi.org/10.2307/3647306

Seberg, O., Petersen, G., 2009. How many loci does it take to DNA barcode a crocus? PLoS ONE 4, 2–7. https://doi.org/10.1371/journal.pone.0004598

Siddiqui, S.A., Islam, Rafiquel, Islam, Rezuanul, Jamal, A.H.M., Parvin, T., Rahman, A., 2017. Chemical composition and antifungal properties of the essential oil and various extracts of *Mikania scandens* (L.) Willd. Arabian Journal of Chemistry 10, S2170–S2174. https://doi.org/10.1016/j.arabjc.2013.07.050

Smith, S.A., Donoghue, M.J., 2008. Rates of molecular evolution are linked to life history in flowering plants. Science 322, 86–89. https://doi.org/10.1126/science.1163197

Soliman, S.S.M., Hamoda, A.M., Soliman, S.S.M., Abou-Hashem, M.M.M., Abouleish, M., Hamoda, A.M., Hamoda, A.M., El-Keblawy, A.A., El-Keblawy, A.A., 2019. Lipophilic metabolites and anatomical acclimatization of *Cleome amblyocarpa* in the Drought and Extra-Water areas of the arid Desert of UAE. Plants 8, 1–12. https://doi.org/10.3390/plants8050132

Spooner, D.M., 2009. DNA barcoding will frequently fail in complicated groups: An example in wild potatoes. American Journal of Botany 96, 1177–1189. https://doi.org/10.3732/ajb.0800246

Stein, E.D., Martinez, M.C., Stiles, S., Miller, P.E., Zakharov, E. V., 2014. Is DNA barcoding actually cheaper and faster than traditional morphological methods: Results from a survey of freshwater bioassessment efforts in the United States? PLoS ONE 9. https://doi.org/10.1371/journal.pone.0095525

Stein, S.E., 1999. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. Journal of the American Society for Mass Spectrometry 10, 770–781. https://doi.org/10.1016/S1044-0305(99)00047-1

Sumner, L.W., Mendes, P., Dixon, R.A., 2003. Plant metabolomics: Large-scale phytochemistry in the functional genomics era. Phytochemistry 62, 817–836. https://doi.org/10.1016/S0031-9422(02)00708-2

Tate, J.A., Simpson, B.B., 2003. Paraphyly of *Tarasa* (Malvaceae) and diverse origins of the polyploid species. Systematic Botany 28, 723–737. https://doi.org/10.1043/02-64.1

Tayade, A.B., Dhar, P., Kumar, J., Sharma, M., Chauhan, R.S., Chaurasia, O.P., Srivastava, R.B., 2013. Chemometric profile of root extracts of *Rhodiola imbricata* Edgew. with hyphenated gas chromatography mass spectrometric technique. PLoS ONE 8. https://doi.org/10.1371/journal.pone.0052797

Trygg, J., Wold, S., 2002. Orthogonal projections to latent structures (O-PLS). Journal of Chemometrics 16, 119–128. https://doi.org/10.1002/cem.695

Uritu, C.M., Mihai, C.T., Stanciu, G.D., Dodi, G., Alexa-Stratulat, T., Luca, A., Leon-

Constantin, M.M., Stefanescu, R., Bild, V., Melnic, S., Tamba, B.I., 2018. Medicinal plants of the family Lamiaceae in pain therapy: A review. Pain Research and Management 2018. https://doi.org/10.1155/2018/7801543

USDA, NRCS, 2021. The Plants Database [WWW Document]. National Plant Data Team. URL https://plants.usda.gov/java/ClassificationServlet?source=display&classid=Linales (accessed 1.20.21).

Van der Kooy, F., Verpoorte, R., Marion Meyer, J.J., 2008. Metabolomic quality control of claimed anti-malarial *Artemisia afra* herbal remedy and *A. afra* and *A. annua* plant extracts. South African Journal of Botany 74, 186–189. https://doi.org/10.1016/j.sajb.2007.10.004

Van Wyk, B., Van Wyk, P., 2013. Field Guide to Trees of Southern Africa, Second edi. ed, Struik publishers. Penguin Random House South Africa.

Viapiana, A., Struck-Lewicka, W., Konieczynski, P., Wesolowski, M., Kaliszan, R., 2016. An approach based on HPLC-fingerprint and chemometrics to quality consistency evaluation of *Matricaria chamomilla* L. Commercial samples. Frontiers in Plant Science 7, 1–11. https://doi.org/10.3389/fpls.2016.01561

Wang, R., Ding, S., Zhao, D., Wang, Z., Wu, J., Hu, X., 2016. Effect of dehydration methods on antioxidant activities, phenolic contents, cyclic nucleotides, and volatiles of jujube fruits. Food Science and Biotechnology 25, 137–143. https://doi.org/10.1007/s10068-016-0021-y

White, D.M., Islam, M.B., Mason-Gamer, R.J., 2019. Phylogenetic inference in section *Archerythroxylum informs* taxonomy, biogeography, and the domestication of coca

(*Erythroxylum* species). American Journal of Botany 106, 154–165. https://doi.org/10.1002/ajb2.1224

Wiklund, S., 2008. Multivariate Data Analysis for Omics [WWW Document]. http://metabolomics.se/courses. URL http://metabolomics.se/Courses/MVA/MVA in Omics_Handouts_Exercises_Solutions_Thu-Fri.pdf (accessed 4.23.20).

Wiklund, S., Johansson, E., Sjöström, L., Mellerowicz, E.J., Edlund, U., Shockcor, J.P., Gottfries, J., Moritz, T., Trygg, J., 2008. Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. Analytical Chemistry 80, 115–122. https://doi.org/10.1021/ac0713510

Wilkie, P., Poulsen, A.D., Harris, D., Forrest, L.L., 2013. The collection and storage of plant material for DNA extraction : The Teabag Method. Gardens' Bulletin Singapore 65, 231–234.

World, S., 1987. Principal component analysis. Chemometrics and Intelligent Laboratory Systems 2, 37–52. https://doi.org/10.1201/b17700-1

Worley, B., Powers, R., 2016. PCA as a predictor of OPLS-DA model reliability. Current Metabolomics 4, 97–103. https://doi.org/10.2174/2213235X04666160613122429.PCA

Yu, J., Wu, X., Liu, C., Newmaster, S., Ragupathy, S., Kress, W.J., 2021. Progress in the use of DNA barcodes in the identification and classification of medicinal plants. Ecotoxicology and Environmental Safety 208. https://doi.org/10.1016/j.ecoenv.2020.111691

Zanolari, B., Guilet, D., Marston, A., Queiroz, E.F., De Queiroz Paulo, M., Hostettmann, K., 2005. Methylpyrrole tropane alkaloids from the bark of *Erythroxylum vacciniifolium*. Journal of Natural Products 68, 1153–1158. https://doi.org/10.1021/np040144h