

Supplementary Material

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

Authors: P.S.F. Alberts* and J.J.M. Meyer

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

The following Supplementary Material is available for this article:

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.

Additional data for downloading from the Figshare (<https://figshare.com/>) data repository:

1) Alberts & Meyer. SA Journal of Botany. 2021.

Full financial cost evaluation and calculations:

<https://doi.org/10.25403/UPresearchdata.16587338>

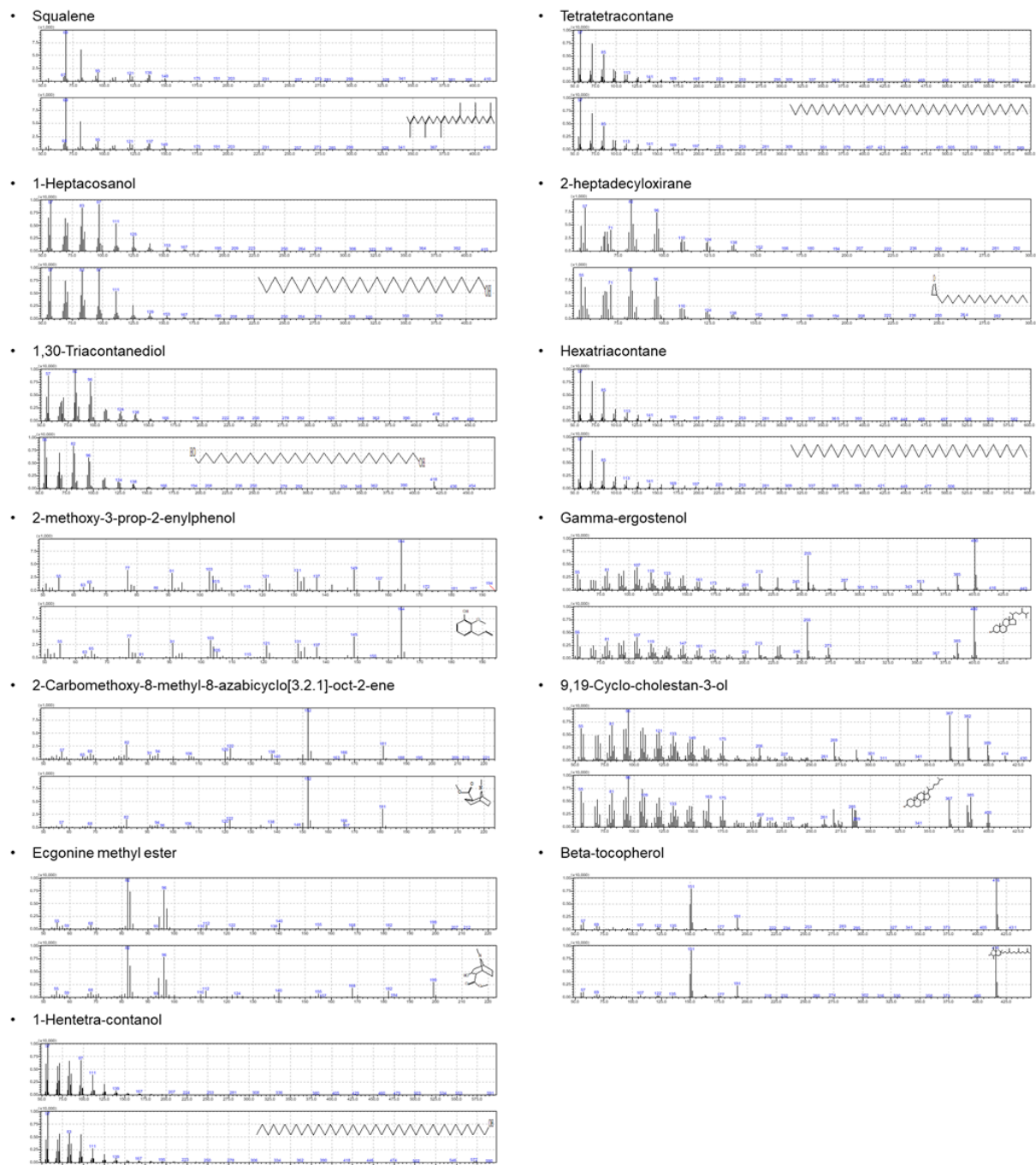


Fig. S. 1. The mass fragmentation patterns of the biomarker (top) compared to the predicted standard compound (bottom) present in the NIST 11 reference database.

Formulate hypothesis & study design
(±2 weeks)

Identification, Collection & Initial Processing
(±273 hours – 35 days)

1. Geographical identification of wild growing plants (1 week).
2. Application of collection permits (1 week).
3. Preparation of items needed for field collections (2 days).
4. Field identification (morphological) and collection of plant material (5 days).
Including; Sorting, processing, and storing leaf material during collection period as well as preparation of Herbarium specimens.
5. Freeze-dry material used for chemotaxonomy and store at 4 °C (3 days).
6. Label and store the DNA barcoding material at -80 °C directly after returning from the fieldtrip (1 hour).
7. Planning and preparing items needed for processing the samples using the three different methods (5 days per method = 10 days).
8. Verify the species collected with Taxonomists at the H.G.W.J. Schweickerdt Herbarium (*Second round of morphological identification and first round of verification*).

Note: Obtaining permits for plant collection can take more than 1 week and is dependent on the response time from the appropriate regulating bodies.

DNA barcoding
(±199 hours – 25 days)

1. Researching and ordering primers needed for PCR & sequencing reactions (1 day).
2. Divide and grind plant material in liquid nitrogen, do not allow plant material to thaw (10 min/ sample = 5 hours).
Autoclaving of apparatus and consumables have already been done.
3. Extract DNA using modified CTAB method and store extracted DNA at -80 °C (2 days/ sample batch of 10 samples = 6 days).
4. Nanodrop and Agarose gel observation of extracted DNA (1 hour/ batch of 10 samples = 3 hours).
Optional: Re-extract DNA from samples with poor quality DNA (Additional 2 days per CTAB batch extraction).
5. Prepare PCR reactions with appropriate barcode primers and preform PCR reactions (2 hours/ batch of 10 samples = 6 hours).
6. Separate PCR reactions on Agarose gel & observe size of fragments (1 hour/ batch of 10 samples = 3 hours).
7. Preform Sequencing PCR reactions with positive PCR products and submit for Sanger cycle sequencing (2 hours/ batch of 10 samples = 6 hours).

Note: Before and after each sequencing preparation, the sample should be cleaned using Sephadex-G50 and MN-Nucleospin columns or another appropriate clean-up method.

8. Import barcode sequences into sequence analysis software (e.g., Geneious Prime) and process as described in the manuscript (1 day).
Including: Aligning, trimming, and verifying the species barcodes using sequence data repositories (e.g., NCBI) if previously submitted sequence data is available for the respective species. This contributes to the genetic identification.
9. Preform associated phylogenetic and statistical analyses as described in the manuscript (2 weeks).

The duration is based on the expertise of the researcher and the research question

Note: Additional steps such as DNA extraction and PCR method optimisations are not included. These steps can have a drastic influence on the time required to complete each method.

Chemotaxonomy

(±544 hours – 68 days, both methods combined:
Separately: ¹H-NMR – 49 days, GC-MS – 50 days)

1. Distillation of the appropriate solvents (2 days).
2. Weighing and extracting the dried plant material (2 hours/ 6 samples = 42 hours).
The same extracts were used for both ¹H-NMR and GC-MS analyses.
3. Drying the extracts to complete dryness as quick and safe as possible (6 hours/ 4 samples = 188 hours).
Store at 4°C (Long term storage at -20°C).
4. Method development of ¹H-NMR and GC-MS (4 hours/ method = 8 hours).
5. Preparing and analysing the ¹H-NMR samples (40 min/ sample = 84 hours).
6. Preparing and analysing the GC-MS samples (45 min/ sample = 94 hours).
Sample preparation should be consistent for both ¹H-NMR, and GC-MS analyses, followed by immediate acquisition and storage of data.
7. Multivariate data analysis and biomarker identification as described in the manuscript (1 week/ method = 28 days).

The duration is based on the expertise of the researcher, the research question and database availability. This contributes to the chemotaxonomic identification.

Note: The time required for full chemotaxonomic identification can vary based on the data acquisition time per sample, method development and the MVDA interpretation.

*1 day = 8 hours, 1 week = 7 days (The time duration was rounded up to the nearest hour or day where appropriate).
The estimated time is based on the samples collected and processed during this study (five individual plants from five species, divided into 125 samples for ¹H-NMR, 125 samples for GC-MS and 25 samples for DNA barcoding).*

Data interpretation, hypothesis testing & discussion

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 site, H.G.W.J. Schweickerdt Herbarium, University of Pretoria (PRU) numbers and the National Centre for Biotechnology Information (NCBI, GenBank) accession numbers of the plant species assessed.

Plant taxon name (Tree #)	Collection sites	PRU numbers	GenBank accession numbers*
<i>Erythroxylum delagoense</i>			
Schinz			
(Tree 1)	Enseleni Nature Reserve	122499	MT476106 (<i>matK</i>); MT476131 (<i>rbcL</i>); MT476167 (<i>trnH-psbA</i>)
(Tree 2)	Enseleni Nature Reserve	122500	MT476107 (<i>matK</i>); MT476132 (<i>rbcL</i>); MT476168 (<i>trnH-psbA</i>);
(Tree 3)	Enseleni Nature Reserve	122501	MT476108 (<i>matK</i>); MT476133 (<i>rbcL</i>); MT476169 (<i>trnH-psbA</i>)
(Tree 4)	Enseleni Nature Reserve	122502	MT476109 (<i>matK</i>); MT476134 (<i>rbcL</i>); MT476170 (<i>trnH-psbA</i>)
(Tree 5)	Enseleni Nature Reserve	122503	MT476110 (<i>matK</i>); MT476135 (<i>rbcL</i>); MT476171 (<i>trnH-psbA</i>)

Table S. 1. (continued)

<i>Erythroxylum emarginatum</i> Thonn.			
(Tree 1)	Mpenjati Nature Reserve	122484	MT476111 (<i>matK</i>); MT476136 (<i>rbcL</i>); MT476172 (<i>trnH-psbA</i>)
(Tree 2)	Mpenjati Nature Reserve	122485	MT476112 (<i>matK</i>); MT476137 (<i>rbcL</i>); MT476173 (<i>trnH-psbA</i>)
(Tree 3)	Mpenjati Nature Reserve	122485	MT476113 (<i>matK</i>); MT476138 (<i>rbcL</i>); MT476174 (<i>trnH-psbA</i>)
(Tree 4)	Mpenjati Nature Reserve	122486	MT476114 (<i>matK</i>); MT476139 (<i>rbcL</i>); MT476175 (<i>trnH-psbA</i>)
(Tree 5)	Mpenjati Nature Reserve	122487	MT476115 (<i>matK</i>); MT476140 (<i>rbcL</i>); MT476176 (<i>trnH-psbA</i>)

Table S. 1. (continued)

<i>Erythroxylum pictum</i> E.Mey. ex Harv. & Sond.			
(Tree 1)	Umtamvuna Nature Reserve	122479	MT476116 (<i>matK</i>); MT476141 (<i>rbcL</i>); MT476177 (<i>trnH-psbA</i>)
(Tree 2)	Umtamvuna Nature Reserve	122480	MT476117 (<i>matK</i>); MT476142 (<i>rbcL</i>); MT476178 (<i>trnH-psbA</i>)
(Tree 3)	Umtamvuna Nature Reserve	122481	MT476118 (<i>matK</i>); MT476143 (<i>rbcL</i>); MT476179 (<i>trnH-psbA</i>)
(Tree 4)	Umtamvuna Nature Reserve	122482	MT476119 (<i>matK</i>); MT476144 (<i>rbcL</i>); MT476180 (<i>trnH-psbA</i>)
(Tree 5)	Umtamvuna Nature Reserve	122483	MT476120 (<i>matK</i>); MT476145 (<i>rbcL</i>); MT476181 (<i>trnH-psbA</i>)

Table S. 1. (continued)

<i>Nectaropetalum capense</i> (Bolus) Stapf & Boodle			
(Tree 1)	Umtamvuna Nature Reserve	122494	MT476121 (<i>matK</i>); MT476146 (<i>rbcL</i>); MT476182 (<i>trnH-psbA</i>)
(Tree 2)	Umtamvuna Nature Reserve	122495	MT476122 (<i>matK</i>); MT476147 (<i>rbcL</i>); MT476183 (<i>trnH-psbA</i>)
(Tree 3)	Umtamvuna Nature Reserve	122495	MT476123 (<i>matK</i>); MT476148 (<i>rbcL</i>); MT476184 (<i>trnH-psbA</i>)
(Tree 4)	Umtamvuna Nature Reserve	122497	MT476124 (<i>matK</i>); MT476149 (<i>rbcL</i>); MT476185 (<i>trnH-psbA</i>)
(Tree 5)	Umtamvuna Nature Reserve	122498	MT476125 (<i>matK</i>); MT476150 (<i>rbcL</i>); MT476186 (<i>trnH-psbA</i>)

Table S. 1. (continued)

<i>Nectaropetalum zuluense</i> (Schönland) Corbishley			
(Tree 1)	Umtamvuna Nature Reserve	122489	MT476126 (<i>matK</i>); MT476151 (<i>rbcL</i>); MT476187 (<i>trnH-psbA</i>)
(Tree 2)	Umtamvuna Nature Reserve	122490	MT476127 (<i>matK</i>); MT476152 (<i>rbcL</i>); MT476188 (<i>trnH-psbA</i>)
(Tree 3)	Umtamvuna Nature Reserve	122491	MT476128 (<i>matK</i>); MT476153 (<i>rbcL</i>); MT476189 (<i>trnH-psbA</i>)
(Tree 4)	Umtamvuna Nature Reserve	122492	MT476129 (<i>matK</i>); MT476154 (<i>rbcL</i>); MT476190 (<i>trnH-psbA</i>)
(Tree 5)	Umtamvuna Nature Reserve	122493	MT476130 (<i>matK</i>); MT476155 (<i>rbcL</i>); MT476191 (<i>trnH-psbA</i>)

*Each GenBank accession number is associated to a specific DNA barcode region i.e., *matK* – maturase K; *rbcL* – ribulose-1,5-bisphosphate carboxylase/ oxygenase large subunit; *trnH-psbA* – intergenic spacer region.

Table S. 2. The DNA primer sequences used in this study.

DNA barcode	Forward primer (5' to 3')	Reverse primer (5' to 3')	Reference
<i>rbcL</i>	ATGTCACCACAAACAGA AAC	TCGCATGTACCTGCAGT AGC	(Kress et al., 2009)
<i>matK</i>	TGAATGGGCTAGTTTCC GGT	CGATTCCGGCAATTGCT CAA	(Primer 3, version 4.1.0)
<i>trnH-psbA</i>	CGCGCATGGTGGATTCA CAATCC	CGAAGCTCCATCTACA AATGG	(Tate and Simpson, 2003)

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

Instrumentation	Average full-service* price for 10 samples at 30 min/sample (South African Rand)**	Average half-service* price for 10 samples at 30 min/sample (South African Rand)**
<i>¹H-NMR</i>	I – ZAR 5124.60	I – ZAR 5124.60
	A – ZAR 10638.00	A – ZAR 8684.10
	C – ZAR 21165.40	C – ZAR 10047.30
<i>GC-MS</i>	I – ZAR 2683.00	I – ZAR 2683.00
	A – ZAR 12195.10	A – ZAR 8938.80
	C – ZAR 17772.08	C – ZAR 12734.00
<i>DNA barcoding (Sanger Cycle Sequencing)</i>	I – ZAR 4860.70	I – ZAR 4860.70
	A – ZAR 8938.80	A – ZAR 7567.10
	C – ZAR 12447.00	C – ZAR 12447.00
<i>HPLC</i>	A – ZAR 5203.30	A – ZAR 3671.60
	C – ZAR 6158.40	C – ZAR 4081.80
<i>LC-MS</i>	A – ZAR 12951.62	A – ZAR 9084.40
	C – ZAR 16781.42	C – ZAR 11924.70
<i>MALDI-MS</i>	A – ZAR 7216.05	A – ZAR 4990.80
	C – ZAR 10267.05	C – ZAR 7102.10
<i>UV-Vis</i>	A – ZAR 3880.60	A – ZAR 3052.10
	C – ZAR 4617.10	C – ZAR 3788.60
<i>ESI-MS</i>	A – ZAR 8354.90	A – ZAR 5151.10
	C – ZAR 11485.00	C – ZAR 8281.20

^A I – In-house rates, A – Academic rates, C – Commercial rates

^B Refer to the link under ‘Additional data for downloading’ for the full financial cost evaluation.

*Full-service prices include the cost of initial data interpretation or report generation by the analysis facility. Half-service represents the cost of data generation without initial interpretation or report generation by the analysis facility.

**The cost of analysis is represented in South African Rand (ZAR) and calculated based on the exchange rates at the time of reporting (USD/ZAR = 14.73, EUR/ZAR = 17.37, GBP/ZAR = 20.60).

Note: These prices are intended for comparative purposes and are subject to change based on the financial exchange rates, availability of instrumentation, sample number, analysis time and the professional relationship between the researcher and analysis facility. We strongly urge any interested persons to contact the analysis facility directly for a detailed cost evaluation based on the research project.