Environmental Advances 5 (2021) 100089

Contents lists available at ScienceDirect

Environmental Advances

journal homepage: www.sciencedirect.com/journal/environmental-advances

A practical study of CITES wood species identification by untargeted DART/QTOF, GC/QTOF and LC/QTOF together with machine learning processes and statistical analysis

Pamela Brunswick[a,*], Daniel Cuthbertson[b], Jeffrey Yan[a], Candice C. Chua[a,†], Isabelle Duchesne[c], Nathalie Isabel[c,1], Philip D. Evans[d], Peter Gasson[e], Geoffrey Kite[e], Joy Bruno[a], Graham van Aggelen[a], Dayue Shang[a,*]

[a] Pacific and Yukon Laboratory for Environmental Testing (PYLET), Science & Technology Branch, Environment and Climate Change Canada, North Vancouver, British Columbia, Canada
[b] Agilent Technologies Inc., Santa Clara, California, United States
[c] Canadian Wood Fibre Centre, Canadian Forest Service, Natural Resources Canada, Quebec, Canada
[d] Faculty of Forestry, Department of Wood Science, University of British Columbia, Vancouver, BC, Canada
[e] Royal Botanic Gardens, Kew, Richmond, Surrey, United Kingdom

ARTICLE INFO

Keywords:
DART
QTOF
Dalbergia
Species
Identification
CITES

ABSTRACT

Illegal logging and trafficking of endangered timber species has attracted the world's major organized crime groups, with associated deforestation and serious social damage. The inability of traditional methodologies and DNA analysis to readily perform wood identification to the species level for monitoring has stimulated research on chemotyping techniques. In this study, simple wood extraction of endangered rosewoods (*Dalbergia* spp), amenable to use in the field, produced colorful hues that were suggestive of wood species. A more definitive study was conducted to develop wood species identification procedures using high-resolution quadrupole time-of-flight (QTOF) mass spectrometers interfaced with liquid chromatography (LC), gas chromatography (GC), and Direct Analysis in Real Time (DART). The time consuming process of extracting "identifying" mass spectral ions for species identification, contentious due to their ubiquitous nature, was supplanted by application of machine learning processes. The unbiased software mining of raw data from multiple analytical batches, followed by statistical Random Forest analysis, enabled discrimination between both anatomically and chemotypically similar *Dalbergia* species. Statistical Principal Component Analysis (PCA) scatterplots with 95% confidence ellipses were visually compelling in showing a differential clustering of *Dalbergia* from other commonly traded and lookalike wood species. The information rich raw data from GC or LC analyses offered a corroborative, legally defensible, and widely available confirmatory tool in the identification of timber species.

1. Introduction

Environmental trading has become a sophisticated web of organized crime involving bribery, corruption, and computer hacking for permits (Nellemann, 2012), with a reported USD 152 billion a year in timber trafficking (INTERPOL, 2020). Worldwide, the illegal logging and trafficking leaves behind social damage, while opening up sensitive areas to disease potential and further climate change (Tollefson, 2020). Internationally, the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) provides a framework around which countries can develop effective laws and enforcement. CITES Appendix I includes species threatened with extinction, Appendix II, those likely to become threatened with extinction without reduced trade, and Appendix III, additional species under special watch. Enforcement of trade surrounding these species is the backbone of the fight against deforestation and promotion of sustainable forestry practices. Consequently, border and port enforcement requires the tools to identify timber and wood products down to the species level listed under

* Corresponding authors.
E-mail addresses: pamela.brunswick@ec.gc.ca (P. Brunswick), dayue.shang@ec.gc.ca (D. Shang).

[†] Candice Chua; Student of the University of British Columbia, Vancouver, BC, Canada

[1] Canadian Forest Service, Natural Resources Canada, Quebec, Canada

https://doi.org/10.1016/j.envadv.2021.100089
Received 2 June 2021; Received in revised form 12 July 2021; Accepted 27 July 2021
Available online 28 July 2021

CITES. Traditional taxonomic approaches and infra-red spectroscopy generally identify to the level of genus rather than species and are no longer solely adequate (Gasson, 2011; Pastore et *al.*, 2011; Dormontt et *al.*, 2015). Portable machine vision systems show potential for rapid screening to genus level but are currently in the early stages of anatomical variability capture (Hermanson and Wiedenhoeft, 2011; Ravindran et *al.*, 2020). The ability of DNA fingerprinting is complicated by limited databases and poor extraction quality from wood products (Meyer and Paulay, 2005; Lowe and Cross, 2011; Jiao et *al.*, 2015), with a more promising alternative being bar coding, constrained again by the need for encompassing the taxonomic diversity of each species (He et *al.*, 2018).

The most promising tools for forensic wood species identification currently rely on chemotyping. Various types of mass spectrometry (MS) have been applied to this end, including Fourier transform ion cyclotron resonance (FTICR-MS) (Zhang et *al.*, 2019), LC/MS (Kite et *al.*, 2010), and GC/MS (Yin et *al.*, 2018; Shang et *al.*, 2020), although the method with most momentum is direct analysis in real time with time-of-flight mass spectrometry (DART/TOF-MS). The U.S. Fish & Wildlife Service Forensics Laboratory (USFWSFL) have led the way in this technique, with their remarkable accomplishment in collating a DART/TOF-MS mass spectral library from thousands of wood species (Lancaster and Espinoza, 2012a; Lancaster and Espinoza, 2012b; Espinoza, et *al.* 2014, Espinoza et al., 2015; McClure et *al.*, 2015; Musah et *al.*, 2015; Evans et *al.*, 2017; Wiemann and Espinoza, 2017; Paredes-Villanueva et *al.*, 2018). One aim of the present study was to determine the compatibility of DART, and associated Mass Mountaineer statistical software (Diablo Analytical), with an alternative quadrupole time-of-flight mass spectrometer (QTOF). A further aim of the study was to determine how other non-dedicated, high-resolution MS instruments, in conjunction with compatible machine learning processes and statistical analysis software, could complement DART analysis by providing recognized and legally defensible procedures for routine wood species identification.

2. Materials and methods

2.1. Materials

Wood specimens used are listed in **Supplementary Information (SI) Table 1** and **SI Table 2**, with sources of the specimens listed in **SI Table 3**. There was limited availability of anatomically verified wood reference specimens available for this study.

Acetic acid (≥98% purity), daidzein, and poly(ethylene glycol) av. Mn 380-420 (PEG400) were supplied by Sigma-Aldrich (Oakville, Ontario, Canada), LC/MS grade acetonitrile and methanol from Fisher Scientific (Ottawa, Ontario) and HPLC grade 2-propanol from Caledon Laboratories (Georgetown, Ontario). Aqueous reagents were prepared in ultra-high purity water (MilliQ Plus).

2.2. Wood sample preparation

Thin slivers (<2 to 3 mm thickness) of heartwood specimens were extracted in 1% v/v formic acid in methanol. The weight of wood was varied by coloration; ~50-100 mg for darker, ~75-150 mg for medium, and ~100-200 mg for lighter colored woods. An initial 2 mL of solution was adjusted depending upon initial coloration, with darker extracts diluted further to 3 – 15 mL total volume. Extraction occurred at room temperature, with optimal recovery after overnight storage. Final samples were vortexed, centrifuged at 4645 x *g* for ~5 min, and the supernatant transferred via glass pipettes to vials for storage at -20 ± 5°C. All containers were glass with Teflon lined lids.

2.3. DART/Ion funnel-QTOF instrument and conditions

A DART-SVP-201 ion source (IonSense, Saugus, MA, USA) was connected via an IonSense Vapur Interface (SVPS-200) to an Agilent 6550 iFunnel Series Quadrupole Mass Spectrometer with Time of Flight detector (QTOF) controlled by MassHunter software. DART conditions were those recommended by the manufacturer (IonSense) for the purpose of wood analysis (**SI Table 4**) and similar to published parameters (McClure, 2015). QTOF calibration was confirmed in electrospray positive mode (ES+) prior to DART connection with intermittent calibration checks using poly(ethylene glycol) (PEG) during sample analysis in positive mode. Background subtracted mass spectra (Agilent MassHunter Qualitative software) were individually converted to .txt format for compatibility with Mass Mountaineer (Diablo Analytical),

2.4. GC/QTOF instrument and conditions

GC/QTOF analysis was performed using an Agilent 7890B GC system interfaced with an Agilent 7250 in EI+ mode controlled by MassHunter software. General operating conditions are summarized in **SI Table 5**. To assist in batch-to-batch alignment, a quality control (QC) *Dalbergia latifolia* extract peak at 268.073 *m/z* was used for retention time locking. Caffeine-d9 was used to confirm system suitability prior to analysis and perfluorotributylamine (PFTBA) mass calibration was included randomly within each analytical sequence.

2.5. LC/QTOF instrument and conditions

LC/QTOF analysis was performed using an Agilent Infinity 1290 LC system interfaced with an Agilent 6550 iFunnel QTOF Mass Spectrometer with an Agilent Jetstream Ion Source (AJS) controlled by MassHunter software. MS detection (**SI Table 6**) employed either ES+ or electrospray negative (ES-) mode ionization and daidzein was used to confirm system performance prior to analysis. Conditions for the reverse phase separation are provided in **SI Table 7** and **SI Table 8**. Following observation of the total ion count chromatography (TIC) profiles in ES+ mode, the elution gradient was optimized for ES- mode scans.

2.6. GC/QTOF and LC/QTOF statistical analysis

Raw QTOF data were imported into Agilent's Unknowns Analysis software, deconvoluted using the SureMass Algorithm, and converted to .cef files for import to Mass Profiler Professional (MPP). MPP performed preliminary alignment, frequency filtering retaining compounds, and ANOVA analysis. For Random Forest analysis of GC/QTOF data only, the MPP selected peaks were exported to MassHunter Quantitative Analysis to allow peak review. Results were reimported into MPP for final filtering, statistical analysis (ANOVA p>0.001), and Random Forest model creation. The predictive accuracy of the Random Forest models was validated internally using Out of Bag Error Estimates for 500 trees and externally by testing the model with 2 naïve specimens of each species that the model had not seen during its creation.

Large LC/QTOF datasets for Random Forest analysis employed Agilent Profinder for recursive feature detection and alignment. Grouped multiple peak entities defined by their mass-to-charge ion ratios, retention time, adducts and peak intensity were then exported to MPP for analysis. In MPP the data was frequency filtered, subjected to ANOVA (p >0.05) and Fold Change (FC>10) analysis. As with the GC/QTOF data, Random Forest models were built and validated internally with Out of Bag Error. External validation of the Random Forest model was performed in the Agilent Classifier program, again with naïve samples. PCA scatter plots were also created to help visualize the output of the algorithms.

3. Results and discussion

3.1. DART/QTOF

An Agilent iFunnel-QTOF (QTOF) was re-purposed for wood species identification in combination with a DART module. The simplicity of the

DART technique, holding a sliver of wood in a stream of heated helium vaporizing wood chemicals directly into the mass spectrometer, is appealing to environmental regulatory laboratories worldwide. A source of strength for the published DART/TOF-MS method is the extensive mass spectral library compiled by the USFWSFL (ForeST Database) and its integration with a commercial statistical software (Mass Mountaineer, Diablo Analytical). USFWSFL has worked with several international agencies to apply their technique against illegal logging (Evans et *al.*, 2017; Paredes-Villanueva et *al.*, 2018).

Dalbergia rosewoods were chosen for this study, based on their CITES I and II listings and the available scattering of published data (Lancaster et *al.*, 2012a; Espinoza et *al.*, 2015; McClure et *al.*, 2015). Our DART/QTOF raw data files required manual processing for compatibility with Mass Mountaineer software but, while initial results were promising (**SI** Figure 1), the mass spectra showed dramatically more response for the major ions in relation to the minor ions compared with the DART/JEOL-TOF-MS ForeST Database (Figure 1).

This was not an issue with certain species but was problematic for others. Repeated attempts to increase the lower mass ions by fragmentation led to decreases in responses and poor spectral matching (**SI Figure 2** and **SI Figure 3**). These differences in ion response ratio suggested that direct application of the available ForeST Database may be limited to a specific hardware configuration, supporting a similar conclusion noted in a Global Timber Tracking Network document (Beeckman et *al.*, 2020).

Considering the world prevalence of GC and LC based mass spectrometers, a way forward would be to consider adding ForeST Database sub-sets to correspond with this prevalent instrumentation. It was further noted that application of the Mass Mountaineer ForeST Database depended upon the selection of DART ions representative of a particular species from collated heat maps. The acceptance or rejection of ions left the process open to individual selection. Future consideration could be given to reduce individual bias in this process. At this point in the current study, the collation of an in-house spectral library was began but was interrupted by consistent carryover materials from the DART interface ceramic tube (instrument specific) reaching the MS (**SI Figure 4**). Together with the inability of the set-up to allow infusion of reference solution for continuous mass correction, the collection of DART/QTOF data was deferred.

3.2. Wood extracts and GC/QTOF

Methanol acidified with 1% v/v formic acid, used to extract heartwood in the present study, was chosen for its direct applicability to both LC and GC analysis. Other researchers have employed alternative matrices, including two-phase chloroform/methanol/water extraction for separation of polar and non-polar metabolites by LC (Creydt et *al.*, 2021) and acidified polar solvents for GC chemotyping (Yin et *al.*, 2018; Shang et *al.*, 2020). In the current study, the chosen matrix resulted in colorful hued extracts (Figure 2), currently being explored by PYLET as a rapid field screening aid for frontline enforcement. Rosewoods can be distinguishable from lookalikes that exhibit little extract coloration, such as *Diospyros* spp. and *Swartzia cubensis* (UNODC, 2016), however, while the color can be suggestive of *Dalbergia* spp., it was not definitive

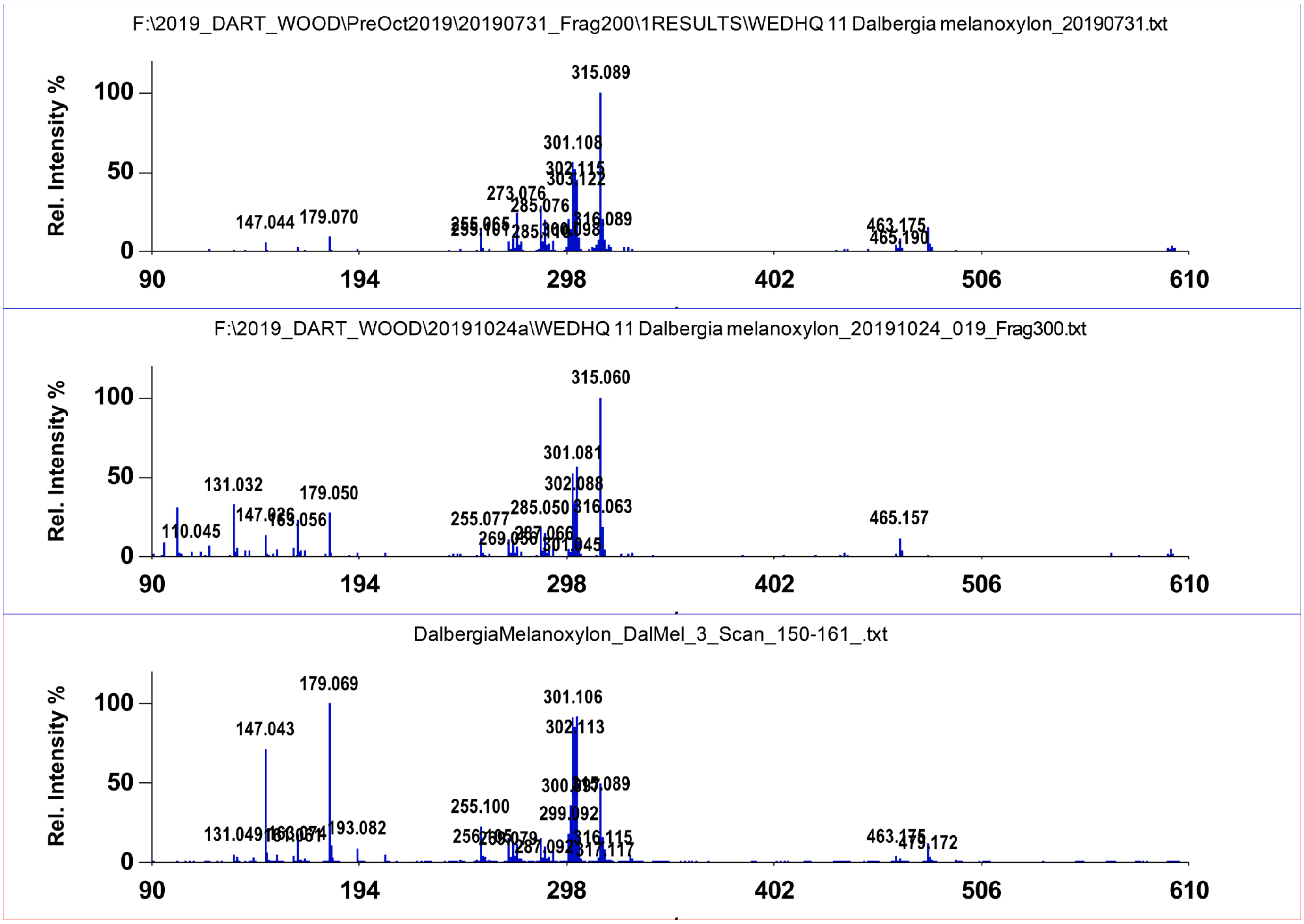


Fig. 1. DART/ Ion Funnel-QTOF for *Dalbergia melanoxylon* using two fragmentor settings (200V top; 300V middle) versus Mass Mountaineer DART/JEOL-TOF-MS database (bottom).

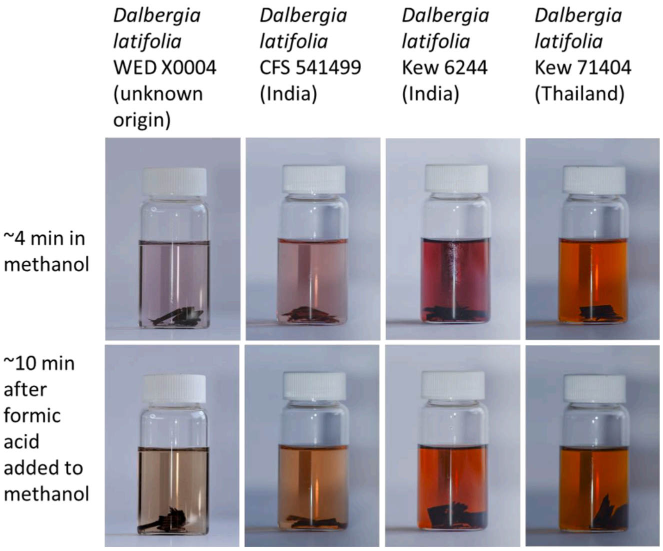

Dalbergia latifolia WED X0004 (unknown origin)
Dalbergia latifolia CFS 541499 (India)
Dalbergia latifolia Kew 6244 (India)
Dalbergia latifolia Kew 71404 (Thailand)
~4 min in methanol
~10 min after formic acid added to methanol

Fig. 2. Example *Dalbergia latifolia* extract colors.

of individual species (**SI Figure 5**).

A confirmatory GC/QTOF and LC/QTOF approach to wood species identification was initiated with the intention of adding the dimension of chromatographic compound separation. Initial studies focused on GC/MS instrumentation, a staple in most analytical laboratories, with its readily repeatable EI+ mass spectra and available NIST database library. A qualitative visual overview of GC/QTOF data showed distinctive total ion chromatography (TIC) for different wood species (Figure 3; **SI Figure 6** through **SI Figure 14**). The close chemical profiles reported between *D. tucurensis* and *D. stevensonii* analyzed by DART/TOF (Espinoza et *al.*, 2015) were not a limitation with the added dimension of chromatographic separation (**SI Figure 8** and **SI Figure 12**).

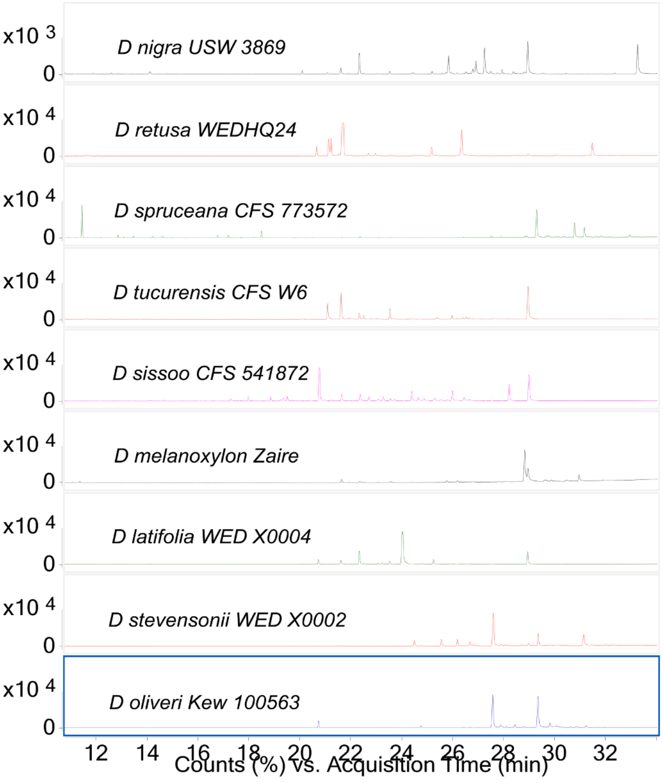


Fig. 3. EI+ GC/QTOF example TIC chromatograms for *D. nigra, D. retusa, D. tucurensis, D. sissoo, D. melanoxylon, D. latifolia, D. stevensonii, D. oliveri,* and *D. spruceana.*

Many *Dalbergia* specimens showed remarkable repeatability of TIC pattern that was immediately indicative of species, eg. *D. latifolia* and *D. sissoo* (**SI Figure 7** and **SI Figure 9**), while others exhibited more variation. For example, *D. nigra* intra-species variation was apparent (**SI Figure 6**), although this did not prevent its visual differentiation from the anatomically similar *D. spruceana* (Figure 3). The diversity, phylogeographic structure and demographic history of *D. nigra* has previously been recognized (Ribeiro et *al.*, 2011). Overall, visual TIC assessment was often informative but not always straightforward, with random peaks caused by natural wood variation, wood processing, and common GC column bleed causing variation (**SI Figure 13**).

Since the CITES I listed *D. nigra* was of particular concern environmentally, specimens of this species and the close lookalike, *D. spruceana*, were reviewed by published identifying ion extraction. For *D. nigra*, the reported dalnigrin identifier (Kite et *al.*, 2010) was likely the peak observed at 298.0836 *m/z* by GC/QTOF and spectral peak formula identification supported this conclusion for all *D. nigra* specimens (**SI Figure 15** and **SI Figure 16**). The anatomically similar *D. spruceana* samples all showed pseudobaptigenin identifier (Kite et *al.*, 2010) confirmed by NIST library search (**SI Figure 17**). *D. spruceana* specimens also showed a distinctive early eluting strong response peak (Rt ~11.5 min) for which the EI+ GC/QTOF showed a [M+] at *m/z* 208.1084 (C12H16O3^{+}). This likely corresponded to an ion observed at 209.12 *m/z* [M^{+}*] or [M+H]$^{+}$ by Lancaster et *al.* (2012a) using DART/TOFMS. A NIST search of the spectrum suggested the compound was elemicin (**SI Figure 17**), corresponding with Proton Nuclear Magnetic Resonance (PMR) MS analysis (Cook et *al.*, 1978) and DART/TOF-MS spectra (Wiemann and Espinoza, 2017). While the authors suggested this compound as an identifier for *D. spruceana*, it is noted that elemicin is a constituent of several plant species, including *Canarium luzonicum* (Mogana and Wiart, 2011), *D. bariensis* (Yang et *al.*, 2015) and *D. latifolia* (Ni et *al.*, 2019). In fact, a number of the wood compounds previously cited as species “identifiers” were observed to be present to variable degrees in other genera and species, hence these compounds would be better termed “indicators”.

Overall, the time consuming and potentially impossible nature of determining specific ”identifier” ions for every species was a daunting prospect, especially considering observed chemotype similarity between some species eg. *D. stevensonii* and *D. oliveri* (Figure 3). This obstacle prompted us to explore the application of data mining, using the readily compatible Agilent Unknowns Analysis and MPP statistical analysis software. An advantage was the unbiased nature of the machine learning process, in comparison to the subjective manual selection of mass ions used by other procedures for wood identification. Deconvolution of complex chromatographic data, followed by frequency filtering and statistical analysis, enabled efficient software selection of the most stable and discriminating entities from the thousands present. Resulting two-dimensional (2D) PCA plots visually showed the variability within a species group and 95% confidence intervals between groups, readily distinguishing between clusters of the two chemotypically similar *D. stevensonii* and *D. oliveri* species (Figure 4).

Variability in GC/QTOF data collated over multiple analysis batches was reduced by the retention time locking feature of the GC. In this way, endangered rosewoods *Dalbergia latifolia* and *Dalbergia oliveri* were screened against a lookalike genus, *Guibourtia* spp., which itself includes three CITES Appendix II listed species (*G. demeusii, G. pellegriniana* & *G. tessmannii*). Visual inspection of TICs readily distinguished between these specimens (**SI Figure 7** , **SI Figure 11, SI Figure 14**) and the result was substantiated by PCA scatterplot (**SI Figure 18**). Similarly, three *Dalbergia* species formed distinct clusters statistically separate from lookalike *Pterocarpus* spp. and *Millettia leucantha* (Figure 4). For comparison with published data (Deklerck et *al.*, 2017), *D. melanoxylon* was successfully differentiated from *Milicia excelsa* and *Dalbergia* lookalike genus *Platymiscium* spp. (Figure 4). Furthermore, despite the wide chemotype variability in *Dalbergia*, a PCA scatterplot was able to

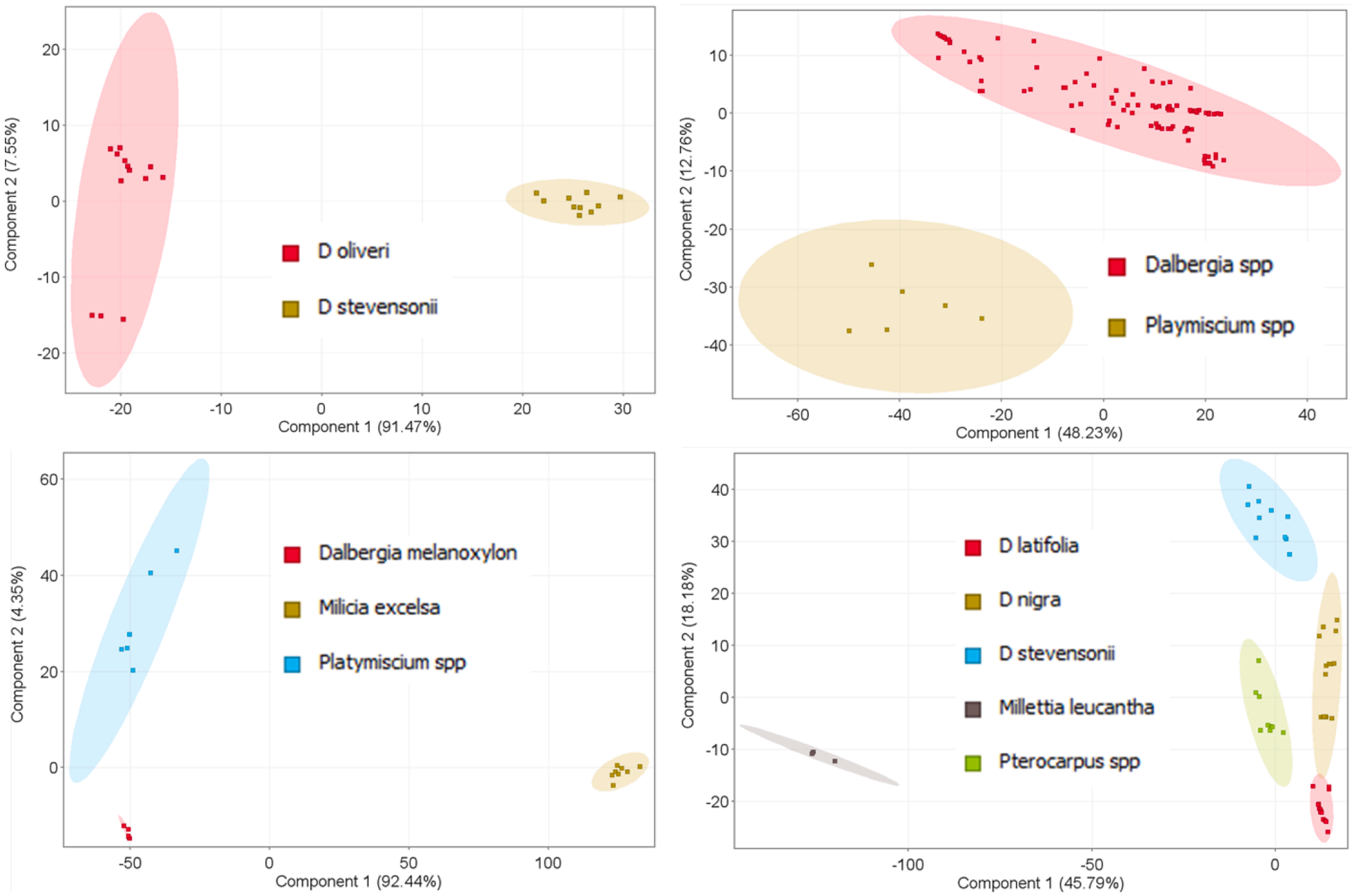
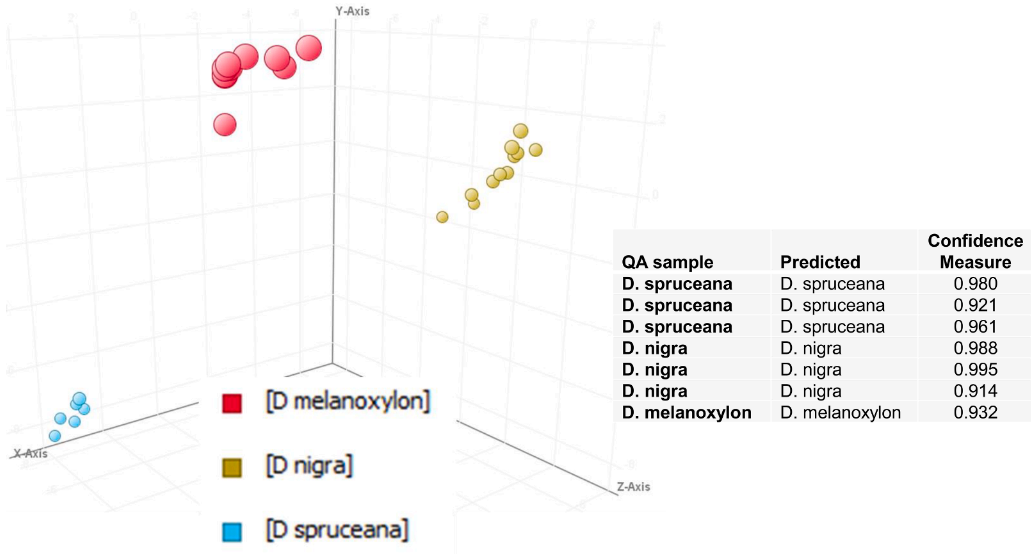


Fig. 4. Principal Component Analysis 2D scatterplots of EI+ GC/QTOF data showing 95% confidence limits for data from distinct sample sources. Top left, *D. stevensonii* and *D. oliveri*; Top right, *D. melanoxylon, Milicia excelsa, Platymiscium* spp.; Bottom left, *Dalbergia* spp. (*latifolia, melanoxylon, nigra, oliveri, retusa, sissoo, stevensonii*) and *Platymiscium* spp. Bottom right, *Dalbergia* spp. (*latifolia, nigra, stevensonii*), *Milettia leucantha* and *Pterocarpus* spp.

collectively group seven *Dalbergia* species, namely *D. latifolia, D. melanoxylon, D. nigra, D. oliveri, D. retusa, D. sissoo*, and *D. stevensonii,* and distinguish the whole group from the lookalike *Platymiscium* spp. (Figure 4).

In consequence of justified restrictions on the trade of the endangered Brazilian rosewood *Dalbergia nigra* (CITES Appendix I), other *Dalbergia* species with similar strength, hardness, color, esthetic appearance, and acoustic properties, have become substitutes. Of these,



QA sample	Predicted	Confidence Measure
D. spruceana	D. spruceana	0.980
D. spruceana	D. spruceana	0.921
D. spruceana	D. spruceana	0.961
D. nigra	D. nigra	0.988
D. nigra	D. nigra	0.995
D. nigra	D. nigra	0.914
D. melanoxylon	D. melanoxylon	0.932

Fig. 5. Partial Least Squares Discrimination preliminary quality assurance testing model limited to *D. nigra, D. melanoxylon*, and *D. spruceana* from distinct samples sources analyzed by EI+ GC/QTOF.

the closest anatomically is *Dalbergia spruceana*. The ability of a Partial Least Squares Discriminant Analysis (PLS-DA) model to distinguish between these closely related species was initially tested using the anatomically similar *D. nigra* and *D. spruceana*, together with an additional lookalike *D. melanoxylon*. Results accurately predicted the correct species of blind quality assurance samples, with a mean 0.954 ±0.03 confidence for *D. spruceana* and 0.966 ±0.045 for *D. nigra* (Figure 5), suggesting that GC/QTOF data offered a robust statistical model for closely related wood species. To address a larger number of species collated from multiple analytical batches, we used a Random Forest decision tree approach to handle the highly complex data and uneven number of replicates per species. Each feature used to build the model was extracted, aligned, and recursively analyzed via Agilent Unknowns Analysis and MassHunter Quantitative analysis to ensure the best statistical reliability of the feature set. To select descriptive features for the modeling approach, only features that passed an ANOVA ($p<0.05$) with all species were chosen. Generally, we found the reliable prediction of *D. spruceana* to be challenging due to the high degree of variability observed in distinct specimens of the species and the limited number of specimens available (SI Table 1). When a Random Forest model was built for eight prevalent *Dalbergia* species, it was able to successfully predict wood species after validation, however, the confidence values for *D. spruceana* were much lower than for other species (SI Table 9). The decreased confidence for *D. spruceana* was also reflected in the boot strapping of the model, after 200 trees the Out of Bag error was 0.06 but *D. spruceana* was at a 0.3 error rate (Figure 6). When the Random Forest model was limited to just *D. spruceana* and a few morphologically similar species (namely *D. melanoxylon* and *D. nigra*), the Out of Bag error and error for each individual species went to zero in less than 50 trees (Figure 6) and validation of the model returned high confidence scores of 0.8 or higher (SI Table 10). Results demonstrated that Random Forest modelling of GC/QTOF data coupled with machine learning methods can be confidently used to predict *Dalbergia* species when sufficient specimens are available.

3.3. Wood identification by LC/QTOF

Wood extracts were further analyzed by Agilent Jetstream ES+ and ES- modes of LC/QTOF, using the same mass spectrometer instrument as for DART analysis. In contrast to GC/QTOF, LC/QTOF provided fewer visual TIC features to readily distinguish specific wood species, both in ES+ (**SI Figure 19, SI Figure 22** to **SI Figure 26**) and ES- modes (**SI Figure 20, SI Figure 21, SI Figure 27** to **SI Figure 35**).

In support of the previous GC/QTOF results confirming "indicator" ions for *D. nigra* and *D. spruceana* specimens, LC/QTOF results for *D. nigra* compared well with the in-depth chemotyping reported by Kite et *al.* (2010) using similar conditions but a different mass spectrometer (**SI Figure 36**). *D. nigra* specimens (CITES Appendix I) exhibited a probable dalnigrin component at 299.0914 *m/z* in ES+ and 297.0768 *m/z* in ES- (**SI Figure 37** and **SI Figure 38**). Caviunin, previously reported using two-dimensional (2D) DART/TOF-MS (Lancaster, 2012a; Wiemann and Espinoza, 2017), was also observed (**SI Figure 39** and **SI Figure 40**). For the anatomically similar *D. spruceana* (CITES Appendix II), pseudobaptigenin was observed at 283.0601 *m/z* in ES+ (**SI Figure 41**) and 281.0453 *m/z* in ES- (**SI Figure 42**). A targeted "Find by Formula" algorithm (Agilent MassHunter Qualitative software) found these compounds to varying degrees in multiple other *Dalbergia* species. However, while results suggested the presence of these compounds and mass spectra were complimentary, further confirmation by fragmentation pattern was not applied due to time constraints. Further, it had been reported that extraction of ions could not be used as a visual indicator of *D. spruceana* without time-consuming individual mass-spectral

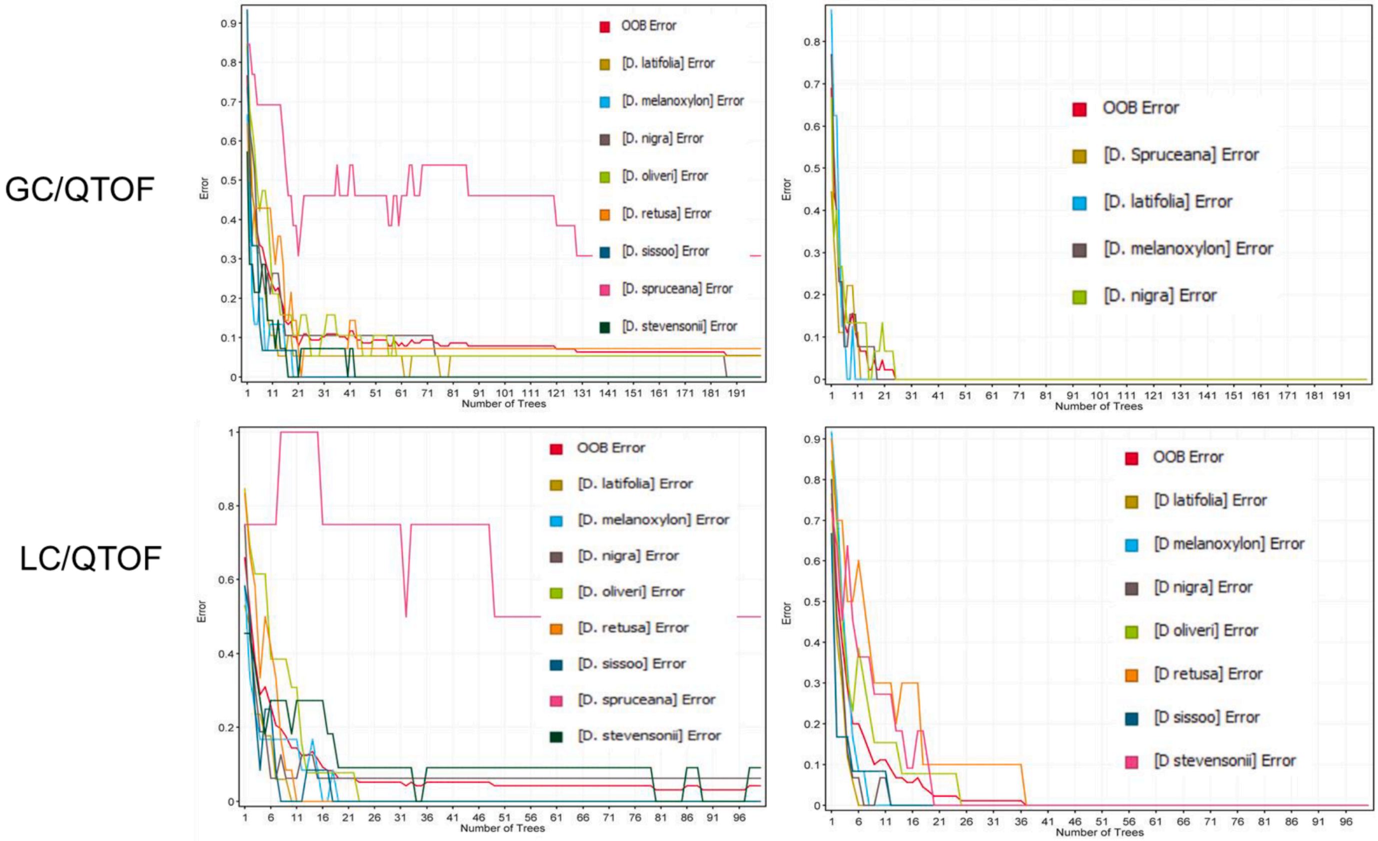


Fig. 6. Out of Bag Error estimates for a Random Forest Model after 200 trees with GC/QTOF data (left) and ES- LC/QTOF (right) from eight prevalent *Dalbergia* species (top) and four prevalent *Dalbergia* including *D. spruceana* (bottom left) or seven prevalent *Dalbergia* excluding *D. spruceana* (bottom right).

assessment to distinguish each from other species (Kite et *al.,* 2010). Thus, conclusions drawn based on visual interpretations alone may not be rigorous enough for a court of law, leaving results open to challenge. Since many of the peaks and ions observed in the chromatography appeared to varying degrees in more than one species, or even genus, the ability of data mining software to sift through the multiple dimensions of information was essential to corroborate visual conclusions. Indeed, a very recent publication has emphasized the potential of similar software procedures to identify plant geographical source (Creydt et *al.,* 2021). With our current LC/QTOF data, application of Agilent Profinder software allowed for untargeted peak detection and alignment of peaks using accurate mass, retention time, and isotope pattern, thus enabling subsequent statistical analysis. Initial tests successfully compared several *Dalbergia* species with another heavily traded precious wood genus, *Pterocarpus* (Figure 7) (Wood Database (The), 2020). Four unconfirmed specimens of *Caesalpinia*, a *Dalbergia* lookalike, were also differentiated from the grouped *Dalbergia* species (Figure 7).

Dalbergia nigra (CITES Appendix I) remains one of the most restricted of the rosewoods, while other *Dalbergia* substitutes include *D. spruceana, D. latifolia, D. stevensonii*, and *D. retusa* (Wood Database (The), 2020), all currently CITES Appendix II listed. Of these species, a close similarity was observed between *D. latifolia* versus *D. sissoo*, and *D. oliveri* versus *D. stevensonii* by both GC/QTOF (Figure 3) and DART/QTOF analysis (**SI Figure 43**). In the current study, PCA and Partial Least Squares Linear Discriminant Analysis (PLS-LDA) statistical analysis of LC/QTOF data provided a clear visual differentiation between these and other closely related *Dalbergia* spp. (Figure 8) (**SI Figure 44, SI Figure 45**). The *D. spruceana* were sourced from three independent xylaria and, while it is recognized that this initially limited number of specimens (4) could not offer a full statistical profile, results demonstrated the potential for the predictive capability of the procedure.

Additionally adding *D. spruceana* and more *Dalbergia* species (*D. melanoxylon, D. sissoo, D. stevensonii,* and *D. oliveri)* together with *D. nigra, D. latifolia*, and *D. retusa* in the PCA model moved some species groups closer with sample groups more diffuse (**SI Figure 46**). This observation may be attributed to individual features being less statistically indicative of a species overall in a large comparison group. Similarly, when a Random Forest Decision tree was built for all species, it confidently predicted the outcome of species during validation prediction, but confidence values were much lower for the *D. spruceana* samples (**SI Table 11**). Additionally, for the tree bootstrapping, the overall Out of Bag error was 0.042 after 100 trees but it remained high at 0.5 for *D. spruceana* (Figure 6). Although the prediction of many species under this scenario remained highly accurate, given our currently limited number of *D. spruceana* samples, an “all-in one” approach may not work for such scenarios. When *D. spruceana* is left out of the model, prediction values remain high (**SI Table 12**), with the Out of Bag error for the Random Forest models at zero after 100 trees (Figure 6). Thus for LC/QTOF data, a two-step approach is recommended, using a larger Random Forest model when many species are compared, together with a secondary tier PLS-DA sub-models for differentiation of fewer species.

In summary, data mining combined with statistical analysis techniques offer a powerful visualization tool and understandable confidence limits for large data sets, but there are some limitations. As with other published DART/JEOL-TOF statistical analysis (Espinoza et *al.*, 2015; McClure et *al.,* 2015; Deklerck et *al.,* 2017), results presented are dependent upon the comparisons made within each statistical data set. Statistical models can become misleading where raw data is poor or, where selective comparison of disparate species groups or inappropriate sampling features have been manually chosen. Considering these limitations are applicable to all statistical MS analysis techniques, our study demonstrated that both the GC/QTOF and LC/QTOF platforms were highly capable of generating profiles in which statistical models could be built and applied for prediction of *Dalbergia* species. The versatility of the machine learning process further enables its application to unit mass resolution data, such as that collected by routine GC/MS instruments. These machine learning processes apply algorithms to the mass spectral data, enabling a more accurate decision making process. Evidence of the software weighting species by neutral mass and retention time can be seen in the Variable Importance and Projection (VIP) score plot for our model (Figure 9). In this step, the software determined how it should appropriately weigh each component in it’s decision making. In fact, machine learning software development was initiated using LC/MS data and has proven invaluable in areas of scientific identification, such as proteomics. There appears to be great potential for these procedures to be adopted by the wood species identification response against the trafficking of endangered species.

4. Conclusion

The current study assessed mass spectrometry procedures in support of wood species identification for enforcement purposes. A DART module paired with a high-resolution Agilent iFunnel-QTOF mass spectrometer produced notably higher response ions for major phytochemicals, while filtering out of the lesser ions in comparison to the ForeST Database DART/JEOL-TOF spectra. The observed statistical differences in ion response ratio did not allow direct application of the ForeST Database, which may be limited to a specific hardware configuration. Future potential remains for compiling a DART mass spectral database with other compatible mass spectrometer instruments.

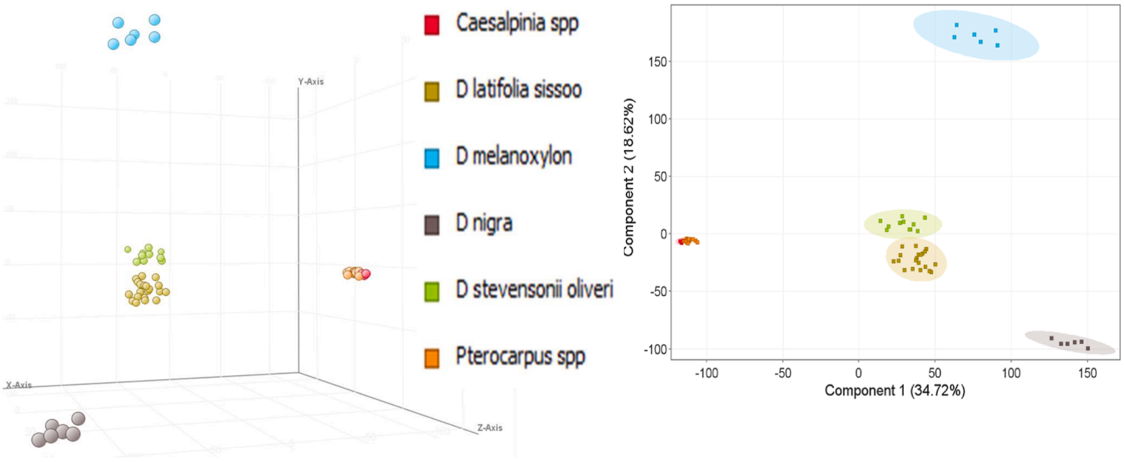


Fig. 7. Principal Component Analysis of ES+ LC/QTOF data for a variety of *Dalbergia* spp. (n=49) and other genus, *Pterocarpus* and *Caesalpinia* from distinct samples sources. It is noted that *D. latifolia* was combined with *D. sissoo*, and *D. stevensonii* combined with *D. oliveri.*

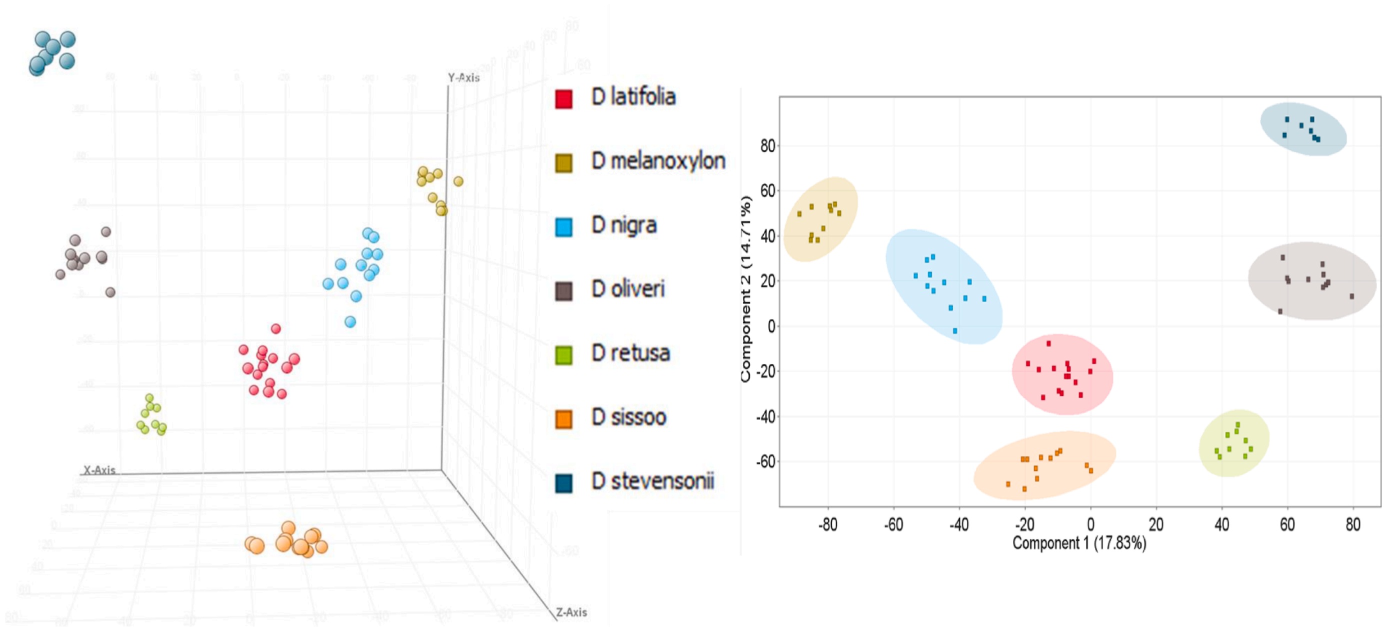


Fig. 8. Principal Component Analysis 3D and scatterplot (showing 95% confidence limits) of seven *Dalbergia* spp *(nigra, melanoxylon, sissoo, latifolia, stevensonii, oliveri,* and *retusa*) from distinct samples sources by ES- LC/QTOF over multiple analytical batches

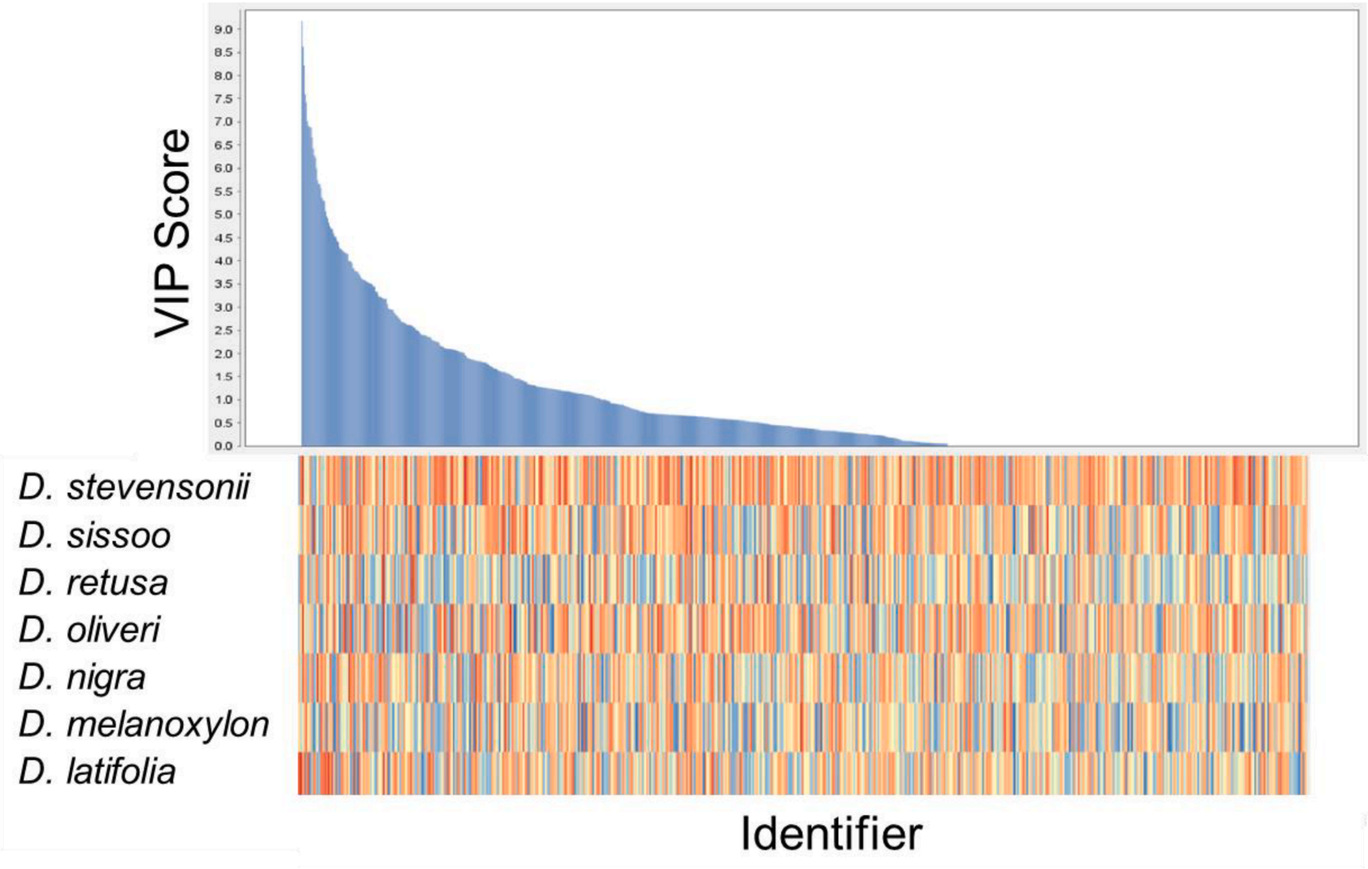


Fig. 9. Variable Importance and Projection (VIP) score plot for Random Forest Model assessing seven *Dalbergia* species by LC/QTOF.

Contemporary GC/MS, with peak retention time locking feature to compensate for intra-batch variability, produced reliable wood species identification results, together with consistent EI+ spectra offering NIST library search capability. While such GC based analysis was lengthier in comparison to DART analysis, the procedure allowed for unattended overnight sequence analysis of hundreds of extracts. Further, the inclusion of chromatographic compound separation offered an additional dimension, with corresponding ability to avoid wood processing contaminants and provide vastly increased minable raw data. Application of machine learning processes significantly supplemented basic visual interpretation of both EI+ GC/QTOF and ES+ or ES- LC/QTOF data. Software algorithms assessed the mass spectral data by accurate unbiased selection of relevant species, enabling statistical software to distinguish between closely related anatomically and chemotypically similar *Dalbergia* species, in addition to their differentiation from other genera. It is anticipated that all procedures, including initial screening by DART/JEOL-TOF and machine vision systems, together with confirmatory analysis by the presented advanced machine learning process procedures of GC/QTOF and LC/QTOF, and indeed basic GC/MS and LC/MS, will form the corroborative basis of future response to illegal logging and wood trafficking.

Author contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Synopsis statement

High-resolution mass spectrometry and machine learning identifies wood species in combat against illegal logging and trafficking of endangered wood species.

Acknowledgment

The authors wish to acknowledge the introductory DART training and encouragement provided by Ed Espinoza, Cady Lancaster, and Pam McClure at the US Fish and Wildlife Science Forensics Laboratory, Oregon, U.S.A.. Further acknowledgement is given to the sourcing of specimens made feasible by the support of Ken Farr, Martine Blais and others of Natural Resources Canada, Quebec, Canada, and by Jean-Francois Dubois and Hector Nieba of Wildlife Enforcement Directorate, Environment and Climate Change Canada, Gatineau, Quebec. Appreciation is given for the generous contribution of specimens made by Prof. Alain Cloutier of Faculty of Forestry, Geography and Geomatics, Department of Forest and Wood Sciences, Université Laval, Quebec, Canada, and the U.S. National Herbarium, Smithsonian Institute, Washington, DC, U.S.A. Special thanks given to Chip (Robert) Cody of JEOL U.S.A, Inc., Danny Rioux of Natural Resources Canada,and Dr. Marcus Kim of Agilent Technologies, for their invaluable support and advice. Last, but not least, we gratefully acknowledge the blessing of supportive colleagues, notably Honoria Kwok, Liane Chow, and Oxana Blajkevitch of Environment and Climate Change Canada (North Vancouver, BC, Canada).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.envadv.2021.100089.

References

Beeckman, H., Blanc-Jolivet, C., Boeschoten, L., Braga, J.W.B., Cabezas, J.A., Chaix, G., Crameri, S., Degen, B., Deklerck, V., Dormontt, E., Espinoza, E., Gasson, P., Haag, V., Helmling, S., Horacek, M., Koch, G., Lancaster, C., Lens, F., Lowe, A., Martínez-Jarquín, S., Nowakowska, J.A., Olbrich, A., Paredes-Villanueva, K., Pastore, T.C.M., Ramananantoandro, T., Razafimahatratra, A.R., Ravindran, P., Rees, G, Soares, L.F., Tysklind, N., Vlam, M., Watkinson, C., Wheeler, E., Winkler, R., Wiedenhoeft, A.C., Zemke, V.Th., Zuidema, P., 2020. Overview of current practices in data analysis for wood identification. A guide for the different timber tracking methods. https://doi.org/10.13140/RG.2.2.21518.79689.

Cook, J.T., Ollis, W.D., Sutherland, I.O., Gottlieb, O.R., 1978. Pterocarpans from *Dalbergia spruceana*. Phytochemistry 17 (8), 1419–1422.

Creydt, M., Ludwig, L., Kohl, M., Fromm, J., Fischer, M., 2021. Wood profiling by non-target high-resolution mass spectrometry: Part 1, Metabolite profiling in *Cedrela* wood for the determination of geographical origin. J. Chom. A. 1641, 461993.

Deklerck, V., Finch, K., Gasson, P., Van den Bulke, J., Van Acker, J., Beeckman, H., Espinoza, E., 2017. Comparison of species Classification models of mass spectrometry data: Kernel Discriminant Analysis vs Random Forest; A case study of Afrormosia (*Pericopsis elata* (Harms) Meeuwen). Rapid Commun. Mass Spectrom. 31, 1582–1588.

Dormontt, E.E., Boner, M., Braun, B., Breulmann, G., Degen, B., Espinoza, E., Gardner, S., Guillery, P., Hermanson, J.C., Koch, G., Leong Lee, S., Kanashiro, M., Rimbawanto, A., Thomas, D., Wiedenhoeft, A.C., Yin, J., Zahnen, J., Lowe, A.J., 2015. Forensic timber identification: It's time to integrate disciplines to combat illegal logging. Biol. Conserv. 191, 790–798.

Espinoza, E.O., Lancaster, C., Kreitals, N.M., Hata, M., Cody, R.B., Blanchette, R.A., 2014. Distinguishing wild from cultivated agarwood (*Aquilaria* spp.) using direct analysis in real time and time of-flight mass spectrometry. Rapid Commun. Mass Spectrom. 28 (3), 281–289.

Espinoza, E.O., Wiemann, M.C., Barajas-Morales, J., Chavarria, G.D., McClure, P.J., 2015. Forensic analysis of CITES-protected *Dalbergia* timber from the Americas. IAWA J. 36, 311–325.

Evans, P.D., Mundo, I.A., Wiemann, M.C., Chavarria, G.D., McClure, P.J., Voin, D., Espinoza, E.O., 2017. Identification of selected CITES-protected Araucariaceae using DART TOFMS. IAWA J. 38 (2), 266–281.

Gasson, P., 2011. How precise can wood identification be? Wood anatomy's role in support of the legal timber trade, especially CITES. IAWA J. 32, 137–154.

He, T., Jiao, L., Yu, M., Guo, J., Jiang, X., Yin, Y., 2018. DNA barcoding authentication for the wood of eight endangered Dalbergia timber species using machine learning approaches. Holzforschung 73 (3). https://doi.org/10.1515/hf-2018-0076.

Hermanson, J.C., Wiedenhoeft, A.C., 2011. A brief review of machine vision in the context of automated wood identification systems. IAWA J. 32 (2), 233–250.

INTERPOL. Forestry crime: targeting the most lucrative of environmental crimes. 14 December (2020). Accessed 17 May 2021 at: https://www.interpol.int/News-and-Events/News/2020/Forestry-crime-targeting-the-most-lucrative-of-environmental-crimes.

Jiao, L., Liu, X., Jiang, X., Yin, Y., 2015. Extraction and amplification of DNA from aged and archaeological *Populus euphratica* wood for species identification. Holzforschung 69 (8), 925–931.

Kite, G.C., Green, P.W.C, Veitch, N.C., Groves, M.C., Gasson, P.E., Simmonds, M.S.J., 2010. Dalnigrin, a neoflavonoid marker for the identification of Brazilian Rosewood (*Dalbergia nigra*) in CITES enforcement. Phytochemistry 71, 1122–1131.

Lancaster, C., Espinoza, E., 2012a. Analysis of select *Dalbergia* and trade timber using direct analysis in real time and time-of-flight mass spectrometry for CITES enforcement. Rapid Commun. Mass Spectrom. 26, 1147–1156.

Lancaster, C., Espinoza, E., 2012b. Evaluating agarwood products for 2-(2-phenylethyl) chromomes using direct analysis in real time time-of-flight mass spectrometry. Rapid Commun. Mass Spectrom. 26, 2649–2656.

Lowe, A., Cross, H.B., 2011. The application of DNA methods to timber tracking and origin verification. IAWA J. 32, 251–262.

McClure, P.J., Chavarria, G.D., Espinoza, E., 2015. Metabolic chemotypes of CITES protected *Dalbergia* timbers from Africa, Madagascar, and Asia. Rapid Commun. Mass Spectrom. 29, 738–788.

Meyer, C.P., Paulay, G., 2005. DNA Barcoding: Error rates based on comprehensive sampling. PLoS Biol. 2 (12), e422. https://doi.org/10.1371/journal.pbio.0030422.

Mogana, R., Wiart, C., 2011. *Canarium* L.: A Phytochemical and Pharmacological Review. J. Pharmacy Res. 4 (8), 2482–2489.

Musah, R.A., Espinoza, E.O., Cody, R.B., Lesiak, A.D., Christensen, E.D., Moore, H.E., Maleknia, S., Drijfhout, F.P., 2015. A high throughput ambient mass spectrometric approach to species identification and classification from chemical fingerprint signatures. Sci. Rep. 5, 11520.

Nellemann, C., 2012. Green carbon, black trade: Illegal logging, tax fraud and laundering in the worlds tropical forests. A rapid response assessment. U. N. Environ. Prog. GRID-Arendal. ISBN: 978-82-7701-102-108.

Ni, C., Jiang, S-C., Chen, J-T., Ge, S-B., Peng, W-X., 2019. Molecules and Indoor Atmosphere Effect of Rosewood: *Dalbergia latifolia*. Ekoloji 28 (108), 27–31.

Paredes-Villanueva, K., Espinoza, E, Ottenburghs, J., Sterken, M.G., Bongers, F., Zuidema, P.A., 2018. Chemical differentiation of Bolivian *Cedrela* species as a tool to trace illegal timber trade. Forest. Int. J. For. Res. 91 (5), 603–613.

Pastore, T.C.M., Braga, J.W.B., Coradin, V.T.R., Magalhães, W.L.E., Okino, E.Y.A., Camargos, J.A.A., de Muñiz, G.I.B., Bressan, O.A., Davrieux, F., 2011. Near infrared spectroscopy (NIRS) as a potential tool for monitoring trade of similar woods: discrimination of true mahogany, cedar, andiroba, and curupixá. Holzforschung. 65, 73–80.

Ravindran, P., Thompson, B.J., Soares, R.K., Wiedenhoeft, A.C., 2020. The XyloTron: Flexible, Open-Source, Image-Based Macroscopic Field Identification of Wood Products. Plant Sci. https://doi.org/10.3389/fpls.2020.01015.

Ribeiro, R.A., Lemos-Filho, J.P., Ramos, A.C.S., Lovato, M.B., 2011. Phylogeography of the endangered rosewood Dalbergia nigra (Fabaceae): insights into the evolutionary history and conservation of the Brazilian Atlantic Forest. Heredity (Edinb). Jan 106 (1), 46–57.

Shang, D., Brunswick, P., Yan, J., Bruno, J., Duchesne, I., Isabel, N., Van Aggelen, G., Kim, M., Evans, P.D., 2020. Chemotyping and identification of protected *Dalbergia* timber using gas chromatography quadrupole time of flight mass spectrometry. J. Chromatogr. A 1615, 460775.

Tollefson, J., 2020. Why deforestation and extinctions make pandemics more likely. News Article. Nature 584, 1750176.

United Nations Office on Drugs and Crime (UNODC), 2016. Best Practice Guide for Forensic Timber Identification. United Nations. Aug.

Wiemann, M., Espinoza, E.O., 2017. Species verification of *Dalbergia nigra* and *Dalbergia spruceana*. Research Paper FPL-RP-690. U.S. Department of Agriculture, Forest ServiceForest Products Laboratory, Madison, WI.

Wood Database (The) accessed 25 August 2020 at: https://www.wood-database.com/brazilian-rosewood/.

Yang, L., Jiang, T., Liu, H., Li, K., 2015. Effects of different drying treatments on preservation of organic compounds in *Dalbergia bariensis* wood. BioResources 10 (4), 7092–7104.

Yin, X, Huang, A., Zhang, S., Liu, R., Ma, F., 2018. Identification of Three *Dalbergia* Species Based on Differences in Extractive Components. Molecules 23, 2163.

Zhang, M., Zhao, G., Guo, J., Wiedenhoeft, A.C., Liu, C.C., Yin, Y., 2019. Timber species identification from chemical fingerprints using direct analysis in real time (DART) coupled to Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS): comparison of wood samples subjected to different treatments. Holzforschung 73 (11).