Quantitative NMR Methodology for the Authentication of Roasted Coffee and Prediction of Blends

Ian W. Burton, Camilo F. Martinez Farina, Subramanyam Ragupathy, Thirugnanasambandam Arunachalam, Steve Newmaster, and Fabrice Berrué*

ABSTRACT: In response to the need from the food industry for new analytical solutions, a fit-for-purpose quantitative 1H NMR methodology was developed to authenticate pure coffee (100% arabica or robusta) as well as predict the percentage of robusta in blends of 292 roasted coffee samples in triplicate. Methanol was chosen as the extraction solvent, which led to the quantitation of 12 coffee constituents: caffeine, trigonelline, 3- and 5-catecholquinic acid, lipids, cafestol, nicotinic acid, N-methylpyridinium, formic acid, acetic acid, kahweol, and 16-O-methylcafestol. To overcome the chemical complexity of the methanolic extract, quantitative analysis was performed using a combination of traditional integration and spectral deconvolution methods. As a result, the proposed methodology provides a systematic methodology and a linear regression model to support the classification of known and unknown roasted coffees and their blends.

KEYWORDS: quantitative NMR, coffee, arabica, robusta, blends, adulteration

INTRODUCTION

Coffee is one of the most consumed beverages worldwide.1 Out of about 500 species of coffee, there are two varieties with the greatest commercial significance, *Coffea arabica* and *Coffea canephora* var robusta, commonly known as arabica and robusta coffees, respectively.1,2 Arabica and robusta coffees are only cultivated in tropical regions mainly centered in South and Central America, East Africa, and Middle-Southeast Asia, with the two varieties differing by their quality and flavor. Soil, climate, altitude, and maturity of the beans are all important factors that affect both the quality and flavor of the final product.3 The fact that robusta varieties are easier to cultivate and that arabica coffee beans are associated with desirable organoleptic properties has implied that robusta coffee has a lower market value than arabica coffee, effectively making arabica coffee the higher-quality product.4,5 However, because of the differing organoleptic properties from various regions and growing environments, it is possible to create valuable coffee blends with specific organoleptic characteristics by mixing coffee beans from different regions. Because these blends are commonplaces in the industry and the price of robusta is lower than that of arabica, there are financial incentives for diluting high-quality arabica beans with the cheaper robusta beans. This practice of adulteration is especially difficult to detect once the coffee beans have been ground and roasted.6 Detection of these adulterations is important to ensure consumer protection and the ability of the food and beverage industry to source high-quality products; to this end, there is an urgent need to develop innovative analytical solutions to ensure the authenticity and quality of raw materials, functional ingredients, and food products.7 Moreover, there are increasing expectations from the consumer that the food industry implements traceability systems to warrant the origin and compositional value of food commodities. One analytical solution that has been applied in the past is DNA barcoding, which has shown to be an effective technique for the authentication of raw products such as meat, fish, crops, or plants.6,7 However, because DNA barcoding is based on the identification and monitoring of unique genetic biomarkers, its efficacy quickly diminishes for extracts, blends, and products whose DNA quality has been damaged during processing steps involving solvent extraction, blending, and heat treatment and in the specific example of coffee, during the roasting of the coffee beans.10 For such products, chemical-based approaches are often preferred, and we recently developed a nuclear magnetic resonance (NMR)-based methodology allowing for the classification of botanicals and dietary supplements based on the unique fingerprints of their chemical content, namely, chemical barcoding.11

Several analytical techniques have been investigated to characterize the chemical composition of coffee using high-performance liquid chromatography (HPLC), gas chromatography (GC), NMR,14–27 near-infrared (NIR),18,19 and infrared (IR) spectroscopy.20,21 These analytical methods were developed to either monitor targeted coffee constituents or capture chemical fingerprints to discriminate coffee samples based on multivariate statistical methodologies.22–25 They take advantage of the fact that coffee contains a broad array of phytochemicals, some of which have been linked to the health
benefits and flavor in coffee. The phytochemicals include caffeine that acts a biostimulant and contributes to coffee bitterness, chlorogenic acids (CQAs), and their derivatives that have antioxidant properties, lipids mainly consisting of triacylglycerols (TAGs), and the diterpenes cafestol, kahweol, and 16-O-methylcafestol (16-OMC) that have been related to increased levels of serum cholesterol. Trigonelline, a pyridine alkaloid, is an N-methylated derivative of nicotinic acid (vitamin B3) that gets degraded during the coffee roasting process into volatile compounds contributing to the flavors and aroma of coffee as well as decarboxylated into N-methylpyridinium, a chemopreventive agent. Melanoids, high-molecular-weight polymeric substances produced by Maillard reaction during roasting, are responsible for the color and aroma of roasted coffee and are related to antioxidant and anti-inflammatory activities. While cafestol is ubiquitously present in both robusta and arabica, its methylated congener (16-OMC) was shown to be distinctly found in robusta and has been proposed as a chemical marker for determining the presence of robusta in adulterated arabica coffees. Consequently, several analytical methods have been validated for the determination of 16-OMC in roasted coffee, including the official method (DIN 10779) published by Deutsches Institut für Normung, but most of these HPLC-based methods are laborious, time-consuming, and expensive.

NMR has shown to be an analytical technique of choice for the study of complex mixtures and food products, and its application to the characterization of coffee has been recently demonstrated using $^1$H NMR spectroscopy. Quantitative $^1$H NMR spectroscopy offers a robust and reproducible approach for the quantitation of chemically diverse analytes present in complex mixtures. Because all small molecules and organic compounds in solution have an equivalent molar response to the detector, the integrations of resonances for each targeted metabolite lead to their quantitation using either an internal or external standard. The quantitation of kahweol and 16-OMC from $^1$H NMR spectra was previously reported by other research groups. However, many of these NMR-based methodologies used chloroform as an NMR solvent, limiting the approach to the analysis of lipophilic compounds in the coffee extracts (i.e., lipids and diterpenes) so that the monitoring of other relevant coffee metabolites is not being accounted for. Although these studies demonstrated the suitability of NMR for detecting and quantifying small amounts of 16-OMC in coffee blends, it is not clear how these results could be leveraged by the industry for the authentication of imported coffee and blends with limited information (compositions, country of origin, cultivars,

Figure 1. Annotated spectra of methanolic extracts showing low-field (bottom) and high-field (top) regions of roasted arabica coffee. Numbers refer to the integration regions in Table 1. Expansion of 7.7–7.5 ppm shows doublets from 3- and 5-CQA that were integrated using deconvolution.
roasting, etc.). Indeed, small amounts of 16-OMC and kahweol were recently detected in some varieties of arabica and robusta coffees, respectively. Because of the natural variation of 16-OMC observed between cultivars and geographical locations, it is a challenge to authenticate a 100% arabica coffee or predict the percentage of robusta in a blend solely based on the amount of 16-OMC.

Consequently, the development of a fit-for-purpose methodology to authenticate pure coffee (100% arabica or robusta) as well as predict the percentage of robusta in blends through quantitative 1H NMR (qHNMR) has been the main focus of this study of 292 coffee samples in triplicate. Methanol was chosen as the extraction solvent, which led to the quantitation of 12 coffee constituents, used to predict mixtures of robusta and arabica, which include the following: caffeine, trigonelline, 3-cafeoylquinic acid (3-CQA) and 5-cafeoylquinic acid (5-CQA), lipids, kahweol, nicotinic acid, N-methylpyridinium, formic acid, acetic acid, kahweol, and 16-OMC.

### MATERIALS AND METHODS

#### Sample Library and Preparation of NMR Samples.

In collaboration with the verification and certification organization Tru-ID and the NHHP Research Alliance at the University of Guelph, 292 roasted coffees and blends were obtained from various industry stakeholders along the supply chain including roasters, processors, and distributors. The library of samples is representative of the industry and is composed of various coffees and blends with different compositions, varieties, quality grades, and countries of origin (e.g., Brazil, Honduras, Colombia, Mexico, Ethiopia, or Vietnam) when this information was made available by the suppliers. Confirmed pure arabica and robusta roasted coffees (Honduran RC-124 and Vietnamese RC-132, respectively) were blended to achieve mixtures of robusta and arabica, which include the following: caffeine, trigonelline, 3-cafeoylquinic acid (3-CQA) and 5-cafeoylquinic acid (5-CQA), lipids, kahweol, nicotinic acid, N-methylpyridinium, formic acid, acetic acid, kahweol, and 16-OMC.

#### NMR Data Acquisition and Processing.

Quantitative 1H NMR spectra were acquired for each aliquot at 700.15 MHz on a Bruker AVANCE III spectrometer equipped with a 5 mm cryogenically cooled inverse detect probe. Spectra were recorded into 64K complex points over the integral ranges for each compound. The deconvolved trigonelline signal was then used to quantitate the percentage of robusta in a blend solely based on the amount of 16-OMC.

Limits of quantitation (LOQs) for these 12 metabolites are listed in Table 1. Integrals were calculated by adding the intensities of the points over the integral ranges for each compound. The deconvolution of the lines of the resonances for 3-CQA, 5-CQA, and 16-OMC was required to separate each line from overlapping interferences, enabling a more precise quantitation of these NMR signals. Deconvolution was performed using the “mdcon” module within Topspin using a pure Lorentzian line shape fitting to frequency and line width.

By obtaining the NMR data under the strictly quantitative conditions described above, the integrated intensities of the resonances of a coffee and its analyte will have the same molar response as the externalstandard of 4 mM caffeine. Differences in reciprocitity between samples can be accounted for using the calibration of the 90° pulse width as a measure of probe damping. The detailed method has been described previously using the relationship for the molar concentration of the analyte C<sub>A</sub>

\[
C_A (\text{mol/L}) = (I_X/I_Y)(P_{90A}/P_{90S})C_S
\]

where I<sub>X</sub> and I<sub>Y</sub> are the integrated intensities of the analyte resonance and external standard resonance, respectively, normalized to the number of protons under that resonance, P<sub>90A</sub> and P<sub>90S</sub> are the calibrated 90° pulse widths in microseconds for the analyte and standard, respectively, and C<sub>S</sub> is the concentration of the internal standard.

The methodology for the quantitation using resonance lines from Lorentz—Gauss deconvolution was similar to an internal standard qHNMR approach where the easy-to-integrate NMR signal of trigonelline at 9.2 ppm is to be both deconvolved and quantitated using the external standard method. The calibrated measurement of the deconvolved trigonelline signal was then used to quantitate the other deconvolved resonances related to 3-CQA, 5-CQA, and 16-OMC. Finally, the molar concentrations were converted and reported as milligrams of analyte per gram of ground coffee for all the coffee metabolites.

#### Extraction Recovery Experiments.

Four samples (RC 55, RC 155, RC 193, and RC 278) were selected based on their wide...
The percent robusta and projection distance were determined by the methodology reported herein.

**RESULTS AND DISCUSSION**

With all metabolites having an equal molar response, NMR is a pertinent quantitative analytical technique for the analysis of complex mixtures and food and beverage ingredients. This study proposes an innovative qHNMR analytical method for the authentication of roasted coffee and predicting blends (percentage of robusta) while quantifying 12 metabolites of relevance to the coffee, 16-OMC, 3- and 5-CQA, caffeine, and trigonelline were calculated as per the methodology described above.

### Table 2. Quantitative Results of Selected Coffee Samples Highlighted in Figure 3 with Values Being Expressed as Milligrams per Gram of Analyte per Gram of Dried Ground Coffee

<table>
<thead>
<tr>
<th>sample</th>
<th>trigonelline</th>
<th>mean S.D.</th>
<th>kabweol</th>
<th>mean S.D.</th>
<th>caffeine</th>
<th>mean S.D.</th>
<th>16-OMC</th>
<th>mean S.D.</th>
<th>percent robusta</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>RC 40</td>
<td>0.165</td>
<td>0.047</td>
<td>3.064</td>
<td>0.189</td>
<td>0.647</td>
<td>0.074</td>
<td>0.001</td>
<td>0.001</td>
<td>0</td>
<td>0.283</td>
</tr>
<tr>
<td>RC 53</td>
<td>0.788</td>
<td>0.031</td>
<td>8.261</td>
<td>0.111</td>
<td>10.997</td>
<td>0.163</td>
<td>0.020</td>
<td>0.007</td>
<td>0</td>
<td>0.241</td>
</tr>
<tr>
<td>RC 62</td>
<td>0.720</td>
<td>0.057</td>
<td>3.757</td>
<td>0.141</td>
<td>5.627</td>
<td>0.206</td>
<td>0.026</td>
<td>0.014</td>
<td>0</td>
<td>0.202</td>
</tr>
<tr>
<td>RC 107</td>
<td>1.136</td>
<td>0.044</td>
<td>3.492</td>
<td>0.101</td>
<td>13.044</td>
<td>0.236</td>
<td>0.316</td>
<td>0.010</td>
<td>39</td>
<td>0.147</td>
</tr>
<tr>
<td>RC 148</td>
<td>0.963</td>
<td>0.104</td>
<td>4.065</td>
<td>0.051</td>
<td>7.549</td>
<td>0.125</td>
<td>0.000</td>
<td>0.000</td>
<td>0</td>
<td>0.255</td>
</tr>
<tr>
<td>RC 164</td>
<td>0.786</td>
<td>0.106</td>
<td>1.979</td>
<td>0.069</td>
<td>9.479</td>
<td>0.545</td>
<td>0.577</td>
<td>0.011</td>
<td>65</td>
<td>0.044</td>
</tr>
<tr>
<td>RC 193</td>
<td>0.648</td>
<td>0.056</td>
<td>3.208</td>
<td>0.092</td>
<td>3.100</td>
<td>0.139</td>
<td>0.003</td>
<td>0.001</td>
<td>0</td>
<td>0.132</td>
</tr>
<tr>
<td>RC 230</td>
<td>0.532</td>
<td>0.058</td>
<td>1.748</td>
<td>0.050</td>
<td>7.880</td>
<td>0.457</td>
<td>0.650</td>
<td>0.017</td>
<td>70</td>
<td>0.304</td>
</tr>
<tr>
<td>RC 232</td>
<td>0.990</td>
<td>0.026</td>
<td>5.570</td>
<td>0.134</td>
<td>5.792</td>
<td>0.151</td>
<td>0.007</td>
<td>0.001</td>
<td>0</td>
<td>0.222</td>
</tr>
<tr>
<td>RC240</td>
<td>0.659</td>
<td>0.056</td>
<td>0.053</td>
<td>0.004</td>
<td>9.659</td>
<td>0.075</td>
<td>0.928</td>
<td>0.010</td>
<td>100</td>
<td>0.216</td>
</tr>
<tr>
<td>RC 253</td>
<td>1.326</td>
<td>0.004</td>
<td>3.665</td>
<td>0.070</td>
<td>6.277</td>
<td>0.046</td>
<td>0.155</td>
<td>0.004</td>
<td>36</td>
<td>0.085</td>
</tr>
<tr>
<td>RC 254</td>
<td>1.177</td>
<td>0.054</td>
<td>2.505</td>
<td>0.074</td>
<td>7.630</td>
<td>0.173</td>
<td>0.446</td>
<td>0.025</td>
<td>56</td>
<td>0.066</td>
</tr>
<tr>
<td>RC 255</td>
<td>1.064</td>
<td>0.043</td>
<td>1.087</td>
<td>0.051</td>
<td>8.455</td>
<td>0.279</td>
<td>0.715</td>
<td>0.002</td>
<td>81</td>
<td>0.150</td>
</tr>
<tr>
<td>RC 267</td>
<td>0.699</td>
<td>0.015</td>
<td>3.748</td>
<td>0.210</td>
<td>8.596</td>
<td>0.055</td>
<td>0.000</td>
<td>0.000</td>
<td>0</td>
<td>0.265</td>
</tr>
<tr>
<td>RC 270</td>
<td>0.679</td>
<td>0.011</td>
<td>0.332</td>
<td>0.006</td>
<td>9.394</td>
<td>0.158</td>
<td>0.467</td>
<td>0.017</td>
<td>100</td>
<td>0.152</td>
</tr>
<tr>
<td>RC 280</td>
<td>0.756</td>
<td>0.012</td>
<td>0.122</td>
<td>0.018</td>
<td>10.318</td>
<td>0.166</td>
<td>0.526</td>
<td>0.013</td>
<td>100</td>
<td>0.072</td>
</tr>
<tr>
<td>RC 286</td>
<td>1.055</td>
<td>0.022</td>
<td>4.942</td>
<td>0.040</td>
<td>8.665</td>
<td>0.138</td>
<td>0.103</td>
<td>0.007</td>
<td>13</td>
<td>0.182</td>
</tr>
<tr>
<td>RC 291</td>
<td>1.298</td>
<td>0.017</td>
<td>4.262</td>
<td>0.038</td>
<td>8.822</td>
<td>0.064</td>
<td>0.166</td>
<td>0.006</td>
<td>25</td>
<td>0.237</td>
</tr>
</tbody>
</table>

The concentration range of caffeine, kabweol, and 16-OMC in order to estimate the extraction efficiency of the proposed methodology when using a one-step extraction (50 mg/mL) and methanol-d$_4$ as a solvent. Five replicates of each coffee sample (5 × 50 mg aliquots) were accurately weighed into glass centrifuge tubes, and 1 mL of methanol-d$_4$ was added. The methanol/coffee mixtures were sonicated for 30 min at room temperature and centrifuged at 1000 rpm for 10 min. The supernatant was then removed, and the extraction procedure was repeated four times to obtain five consecutive extracts for each coffee replicate. The supernatants (600 μL) of each sample were transferred into NMR tubes for qNMR analysis, and the concentrations of kabweol, 16-OMC, 3- and 5-CQA, caffeine, and trigonelline were calculated as per the methodology described above.

**Table 1.** The concentration range of caffeine, kabweol, and 16-OMC in order to estimate the extraction efficiency of the proposed methodology when using a one-step extraction (50 mg/mL) and methanol-d$_4$ as a solvent. Five replicates of each coffee sample (5 × 50 mg aliquots) were accurately weighed into glass centrifuge tubes, and 1 mL of methanol-d$_4$ was added. The methanol/coffee mixtures were sonicated for 30 min at room temperature and centrifuged at 1000 rpm for 10 min. The supernatant was then removed, and the extraction procedure was repeated four times to obtain five consecutive extracts for each coffee replicate. The supernatants (600 μL) of each sample were transferred into NMR tubes for qNMR analysis, and the concentrations of kabweol, 16-OMC, 3- and 5-CQA, caffeine, and trigonelline were calculated as per the methodology described above.

**The percent robusta and projection distance were determined by the methodology reported herein.

External standard qNMR methodology by recording NMR data on a 4 mM caffeine solution for each batch of samples. As noted in Table 1, the deconvolution of the aromatic region of the caffeoylquinic acids (7.50–7.70 ppm) was required for the quantitation of the closely related chemical structures 3-CQA and 5-CQA in the presence of overlapping resonances from other CQAs. Similarly, the methoxy protons of 16-OMC at 3.17 ppm were deconvolved from other unidentified metabolites, improving the limit of quantitation and the accuracy of the quantitation results. One unique aspect of this methodology is in the use of trigonelline as a calibrant for the results of the deconvolution. The trigonelline resonance at 9.21 ppm is easily integrated, and the molar concentration of trigonelline in any sample is determined using the external standard method. The trigonelline can then be used as an internal standard against the areas of resonances determined through deconvolution which in the Topspin software are relative only within the current spectrum. The terminal methyl attributed to lipids (0.932–0.885 ppm) is only used to estimate the amount of total fatty acids, and the concentration expressed in milligrams per gram was calculated based on the molecular weight of stearic acid. This calculation provides a rough estimate of the lipid content as the multiplet resonance at 5.274 ppm suggests that a significant amount of fatty acids is the building block of triacylglycerol lipids. Moreover, the resonance at 5.35 ppm also indicates the presence of vinyl protons from unsaturated fatty acids. The concentration of each coffee metabolite obtained for the 292 coffee samples is reported in the Supporting Information (Tables S2 and S3), and a subset of the results is provided in Table 2 to illustrate the methodology for authenticating coffee and predicting blends.

### Extraction Efficiency

The use of methanol-d$_4$ as an extraction solvent yields a wider range of chemically diverse molecules than can typically be obtained using either deuterated water (D$_2$O) or chloroform-d. This is at the expense of absolute quantitation of lipophilic compounds (e.g., fatty acid esters of 16-OMC and triglycerides), which have a...
lower solubility in methanol. Consequently, five consecutive methanol extractions were carried out on four coffee samples RC 55, RC 155, RC 193, and RC 278 in five replicates each in order to assess the extraction efficiency. The results (see Figure S4 in the Supporting Information) showed that greater than 98% of kahweol and 16-OMC were recovered after four serial extractions in methanol, with 80% of the total kahweol and 72% of the total 16-OMC being recovered in the first extraction. The replicate data confirmed that a one-step extraction was suitable to reproducibly quantify the concen-
tration of kahweol and 16-OMC in a methanolic extract with comparable extraction efficiencies (70–80%). With any one-step extraction methodology, it is expected that only partial amounts of the metabolites are being recovered in the extract and correction factors have to be applied to report the full amount of each constituent in coffee. In this study, all the concentrations were expressed in milligrams per gram of dried biomass and represent the amount of coffee constituent in a methanolic extract at 50 mg/mL. Moreover, the proposed prediction model determines the percentage of robusta in blends based on the relative abundance of 16-OMC and kahweol rather than their individual concentrations. It should be noted that the 16-OMC concentrations observed in robusta coffees in this study are lower than the values reported by Finotello et al. for other 100% arabica coffees (1.2–2.2 mg/g). This could be explained by the partial solubility of 16-OMC in methanol, especially its fatty acid ester forms, as no saponification step was included in this method. Extraction efficiencies were also calculated for other coffee constituents. Caffeine, 3- and 5-CQA, and trigonelline showed slightly lower amounts recovered in the first extraction (67, 52, 55, and 53%, respectively) when compared to kahweol and 16-OMC. For comparison, caffeine is not quantifiable in chloroform extracts, while trigonelline, nicotinic acid, CQAs, and other organic acids are not detected, as indicated in the Supporting Information, S7 and S8.

Linearity and Reproducibility. With kahweol and 16-OMC being distinct chemical markers associated with arabica and robusta, respectively, it was hypothesized that the quantitation of these two metabolites and their correlation could be leveraged for predicting blends based on a simple linear regression. To demonstrate this hypothesis, the reproducibility and linearity of the methodology were evaluated by analyzing known blends of arabica and robusta roasted coffees (Honduran RC-124 and Vietnamese RC-132, respectively) with the following percentages of robusta: 0, 10, 20, 30, 40, 50, 60, 80, and 100%. qHNMR analyses were carried out on each blend in triplicate. With the y-intercept representing 100% robusta and the x-intercept representing 100% arabica in Figure 2, the resulting linear regression model \( y = -0.2281x + 0.8272 \) showed a good fit with the experimental data with an R-squared value of 0.99. It is also worthwhile noticing that good reproducibility was observed between replicates and the larger discrepancies are most likely due to the heterogeneity of the samples during mixing of the blends. Interestingly, the data suggested that Vietnamese robusta (RC-132) contains a measurable amount of kahweol (0.19 mg/g), which is in accordance with observations previously reported in the literature. It is unclear if the trace amounts of kahweol in robusta samples are native or due to cross-contamination during harvest and processing, but this variance was accounted for in the validation of the prediction model for authenticating roasted coffees (100% arabica or 100% robusta) and the percentage of robusta for blends.

Authentication of Roasted Coffees and Prediction of Blends. Based on the quantitative results obtained for 292 roasted coffee samples, a scatter plot (Figure 3) was created with the entire data set by correlating the concentrations in kahweol and 16-OMC (the x-axis and y-axis, respectively). The mean and standard deviations (S.D.s) were calculated based on the triplicate analysis of each sample. The correlated concentrations formed an overall trend in the data set following the regression line \( y = -0.1099x + 0.6217 \) with an R-squared value of 0.72. Because of the size of the data set, a subset of 18 representative coffee samples was selected along the trend line to support the discussion and illustrate the methodology. The key quantitative results for these selected samples are summarized in Table 2, and their locations and their coordinates have been highlighted in the scatter plot (Figure 3). To predict the percentage of robusta in coffee blends, it was important to first define two concentration thresholds to classify coffee samples considered as 100% arabica (low 16-OMC) and robusta (low kahweol). As previously mentioned, small amounts of 16-OMC and kahweol have been reported in arabica and robusta, respectively. Similar results were obtained in this study, as indicated by the wide distributions of samples along both x- and y- axes. Consequently, the threshold values in 16-OMC and kahweol concentrations were defined as 0.04 and 0.35 mg/g, respectively, based on the LOQ values of both 16-OMC and kahweol as well as the distribution of the 292 samples along the x-axis and the linear regression lines in the model (Figure 3). It should be noted that the LOQ values for 16-OMC and kahweol (0.026 and 0.252 mg/g, respectively) are lower than the selected thresholds, suggesting that the authentication of coffee (100% arabica and 100% robusta) could be achieved using lower concentration values for more specific and defined sets of coffee samples. For reference, the selected concentration thresholds correspond to a level of contamination estimated at approximately 6% by the linear regression model. By applying this methodology, the coffee samples RC40, RC62, RC148, RC193, RC232, and RC267 located along the x-axis displayed a concentration of 16-OMC below 0.04 mg/g (Table 2) and were classified as 100% arabica. Similarly, the coffee samples RC 240, RC270, and RC280 were identified as 100% robusta with a low amount of kahweol detected (0.05, 0.33, and 0.12 mg/g, respectively).

The rest of the samples were classified as blends, and the amount of robusta in each sample was estimated using the projection of each point on the regression line and reported as the distance along this line from the x-intercept and y-intercept. Although the composition of the blends was not made available for most of these samples, the proposed methodology and regression line \( y = -0.1099x + 0.6217 \) appeared to provide good estimates for blends whose compositional information was provided. As examples, the coffee samples RC253–RC255 were labeled as Brazil Arabica/Vietnam Robusta blends with proportions of arabica/robusta of 80/20, 40/60, and 20/80, respectively. Their projections on the trend line led to estimated amounts of robusta of 36, 56, and 81%, respectively, for the three samples (Table 2). When projection to the trend line provided negative values, the percentage in robusta was reported as 0 or 100%, and the samples were qualified as atypical. For example, the coffees RC 53 and RC 240 were considered as outliers in the given regression model and displayed the highest concentrations in kahweol (8.261 mg/g) and 16-OMC (0.928 mg/g) in this study, respectively.

As a key finding in this study, the classification of coffee samples and the prediction of blend compositions can be achieved without prior information on the coffee samples by using a systematic approach utilizing the linear regression model \( y = -0.1099x + 0.6217 \) and the concentration thresholds defined in this article. Moreover, the addition of more samples in the future will help to further refine the model by studying other cultivars or coffee quality grades as well as...
Figure 4. Normal distribution of concentrations of 16-OMC (right) and kahweol (left) for samples that were identified as either primarily robusta or primarily arabica. Vertical axes are arbitrary units related to the probability density function for each distribution. Plots have been scaled so that the arabica and robusta distribution curves are represented at the same height with the vertical axis on the right corresponding to the distribution of Arabica samples. Concentrations are expressed as milligrams of analyte per gram of dried coffee.

diversifying the library with other geographical locations. Having said this, it is not anticipated that the linear regression line should drastically shift because the model relies on the analysis of 292 diverse coffee samples. Consequently, the predicted percentages of robusta, reported in this study, would remain the same for most of the blends at the exception of few outliers.

As a systematic approach is used for the determination of the blend composition, it was also important to report the results with a metric representing the confidence level in the predicted results. For this, the measurement of the projection distance (D) between the data point and the trend line was found to be suitable. It provided a quantifiable metric to identify outliers in the linear regression model, rapidly highlighting the samples and percent robusta estimates that should be treated more cautiously. Looking at Figure 3 and the distance values listed in Table 2, the coffee blends RC164, RC230, and RC255 would be considered as atypical coffees in the given model with greater distance values (0.164, 0.217, and 0.213, respectively) than RC107, RC286, and RC291 (0.072, 0.057, and 0.023, respectively). The latter coffee blends are therefore in better agreement with the model, with the prediction of blend composition being more robust.

Normal distribution curves were created for kahweol, 16-OMC, caffeine, 3-CQA, trigonelline, and total lipids using samples that were classified as 100% arabica and 100% robusta based on the defined thresholds. The samples were assumed to be normally distributed, and the distribution curve was calculated from the mean and S.D. of the robusta and arabica sample populations. The concentrations of the analytes were then superposed on the calculated curves. The distribution curves for kahweol and 16-OMC are distinct based on the respective pure cultivar, as expected (Figure 4). Other analytes, notably caffeine and lipids, show a partial separation of distributions, whereas trigonelline and 3-CQA display no difference between the distributions (Figure S5). In the robusta samples, the concentration of 16-OMC ranged from 0.53 to 0.93 mg/g with a mean of 0.69 ± 0.14 mg/g, and in the arabica samples, the concentration of kahweol ranged from 3.08 to 8.26 mg/g with a mean of 4.88 ± 1.21 mg/g (errors are ± one S.D.). The difference in distribution between kahweol and 16-OMC in arabica and robusta coffees is the basis for this model. In addition to natural variations, roasting and processing might be other factors responsible for the large distribution of kahweol observed in the data set. Seemly, the decaffeinated coffee RC40 with a measured caffeine content of 0.647 mg/g and RC267 labeled as vanilla-flavored coffee showed concentrations of kahweol significantly lower than average with concentrations of 3.060 mg/g (an S.D. of 0.189) and 3.748 mg/g (an S.D. of 0.210), respectively.

Even though the distributions of 16-OMC and kahweol are quite broad over all geographical varieties, when considering a large number of samples, such as that in this study, the regression trend line still gives reliable results. The study suggests that the variability of the content for both 16-OMC and kahweol in 100% robusta and 100% arabica, respectively, is the main limiting factor for the development of any analytical techniques leveraging these two chemical markers for the predictions of coffee blends. However, correlating the concentrations of 16-OMC to kahweol has added to the robustness and reliability of the proposed methodology over solely using the concentration of 16-OMC as reported in the DIN method and other NMR methodologies. Recording NMR data on the methanolic extract rather than chloroform has also enabled the analysis of a wider range of coffee constituents including the 12 compounds reported in this study, therefore providing supplementary and valuable information to the coffee industry for the same analytical cost. In general, NMR data recorded on methanolic extracts have shown to be highly suitable for untargeted analysis, and preliminary results indicated that multivariate statistical analysis could further discriminate coffee samples identified as 100% robusta and 100% arabica coffee based on their chemical profiles.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jafc.0c06239.

Concentrations of trigonelline, cafestol, kahweol, caffeine, and 16-OMC for all 292 coffee samples; concentrations of total lipids, 5-CQA, 3-CQA, nicotinic acid, N-methylpyridinium, formic acid, and acetic acid for all 292 coffee samples; structures of the major compounds quantified in coffee; results of the extraction efficiency experiment showing cumulative concentrations of trigonelline, 5-CQA, 3-CQA, kahweol, caffeine, and 16-OMC; normal distribution curves for trigonelline, caffeine, 3-CQA, and total lipids; expansion of the

https://dx.doi.org/10.1021/acs.jafc.0c06239
J. Agric. Food Chem. 2020, 68, 14643−14651
1H-13C HSQC-TOCSY spectrum of the sample RC-160 with 16-OMC methoxy correlation from 3.169 to 49.6 ppm; comparison of the chloroform extract and methanol extract of the sample RC 155 highlighting the region from 6.5 to 3.0 ppm; and comparison of the chloroform extract and methanol extract of the sample RC 155 highlighting the region from 10 to 5.5 ppm (PDF).

### AUTHOR INFORMATION

**Corresponding Author**

Fabrice Berrue — Aquatic and Crop Resources Development Research Center, National Research Council of Canada, Halifax, Nova Scotia B3H 3Z1, Canada; orcid.org/0000-0003-2389-4388; Phone: 1-902-426-0636; Email: Fabrice.Berrue@nrc-cnrc.gc.ca

**Authors**

Ian W. Burton — Aquatic and Crop Resources Development Research Center, National Research Council of Canada, Halifax, Nova Scotia B3H 3Z1, Canada; orcid.org/0000-0002-9387-0595

Camilo F. Martinez Farina — Aquatic and Crop Resources Development Research Center, National Research Council of Canada, Halifax, Nova Scotia B3H 3Z1, Canada

Subramaniam Ragupathy — NHP Research Alliance, College of Biological Sciences, University of Guelph, Guelph, Ontario N1G 4T2, Canada

Thirunagasambandam Arunachalam — NHP Research Alliance, College of Biological Sciences, University of Guelph, Guelph, Ontario N1G 4T2, Canada

Steve Newmaster — NHP Research Alliance, College of Biological Sciences, University of Guelph, Guelph, Ontario N1G 4T2, Canada

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jafc.0c06239

**Author Contributions**

I.W.B. and C.F.M.F. contributed equally. The manuscript was written through contributions of all authors.

**Notes**

The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

The authors would like to thank Tru-ID for providing the samples free of charge.

**REFERENCES**


