DOI:10.1021/jf8037117

JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY ARTICLE

Botanical and Geographical Characterization of Green Coffee (*Coffea arabica* and *Coffea canephora*): Chemometric Evaluation of Phenolic and Methylxanthine Contents

Rosa M. Alonso-Salces,*,[†] Francesca Serra,[†] Fabiano Reniero,[†]
and Károly Héberger[†,§]

[†] European Commission - Joint Research Centre, Institute for Health and Consumer Protection, Physical and Chemical Exposure Unit, Via Fermi 2, I-21020 Ispra, Italy and [§] Chemical Research Center, Hungarian Academy of Sciences, P.O. Box 17, H-1525 Budapest, Hungary

Green coffee beans of the two main commercial coffee varieties, *Coffea arabica* (Arabica) and *Coffea canephora* (Robusta), from the major growing regions of America, Africa, Asia, and Oceania were studied. The contents of chlorogenic acids, cinnamoyl amides, cinnamoyl glycosides, free phenolic acids, and methylxanthines of green coffee beans were analyzed by liquid chromatography coupled with UV spectrophotometry to determine their botanical and geographical origins. The analysis of caffeic acid, 3-feruloylquinic acid, 5-feruloylquinic acid, 4-feruloylquinic acid, 3,4-dicaffeoylquinic acid, 3-caffeoyl-5-feruloylquinic acid, 3-caffeoyl-4-feruloylquinic acid, 3-*p*-coumaroyl-4-caffeoylquinic acid, 3-caffeoyl-4-dimethoxycinnamoylquinic acid, 3-caffeoyl-5-dimethoxycinnamoylquinic acid, *p*-coumaroyl-*N*-tryptophan, feruloyl-*N*-tryptophan, caffeoyl-*N*-tryptophan, and caffeine enabled the unequivocal botanical characterization of green coffee beans. Moreover, some free phenolic acids and cinnamate conjugates of green coffee beans showed great potential as means for the geographical characterization of coffee. Thus, *p*-coumaroyl-*N*-tyrosine, caffeoyl-*N*-phenylalanine, caffeoyl-*N*-tyrosine, 3-dimethoxycinnamoyl-5-feruloylquinic acid, and dimethoxycinnamic acid were found to be characteristic markers for Ugandan Robusta green coffee beans. Multivariate data analysis of the phenolic and methylxanthine profiles provided preliminary results that allowed showing their potential for the determination of the geographical origin of green coffees. Linear discriminant analysis (LDA) and partial least-squares discriminant analysis (PLS-DA) provided classification models that correctly identified all authentic Robusta green coffee beans from Cameroon and Vietnam and 94% of those from Indonesia. Moreover, PLS-DA afforded independent models for Robusta samples from these three countries with sensitivities and specificities of classifications close to 100% and for Arabica samples from America and Africa with sensitivities of 86 and 70% and specificities to the other class of 90 and 97%, respectively.

KEYWORDS: Chlorogenic acid; cinnamoyl amide; cinnamoyl glycoside; methylxanthine; coffee; pattern recognition

INTRODUCTION

Coffee is one of the most important food commodities both for producers, that is, countries in the tropical and subtropical areas with coffee as their main agricultural export product, and for manufacturers, which are mainly located in Europe and North America, where coffee is roasted, mixed, and packed. Due to its large diffusion and high market value, it is not unusual that coffee is subject to adulteration throughout its production chain. One of the most common forms of fraud concerns the mislabeling of the product to conceal the true origin of the coffee (botanical and/or geographical origin). The two major coffee species consumed worldwide are *Coffea arabica* (Arabica) and *Coffea canephora* (Robusta). Arabica coffee is considered to be superior to Robusta due to its organoleptic properties, and it is therefore more expensive. Robusta coffee has been characterized as a neutral coffee, weak-flavored, and occasionally with a strong and pronounced bitterness (*1*), whereas Arabica coffee is a higher priced, milder, fruitier, and acidulous beverage (*2*). There are also certain prestigious coffee-growing areas ("terroir"). For instance, Arabica coffees produced in Central America are traditionally the most highly appreciated, whereas Brazilian Arabica coffee is considered to be of lower quality, due to the methods of harvesting (strip-picking) and processing practices used in that country (*3*). Therefore, the detection of fraud concerning the geographical origin of coffees at regional or

*Author to whom correspondence should be addressed (e-mail rosa-maria.alonso-salces@jrc.it; telephone +39-0332785986; fax +39-0332789303).

Published on Web 3/19/2009

national level, as well as the adulteration and mislabeling at the botanical level, are very relevant with special regard to socioeconomic issues.

Several attempts have been made to develop analytical procedures for the botanical characterization of green coffee beans: chlorogenic acids, phenolic acids, and total polyphenols, together with other chemical descriptors, such as metals and other elements (N, P, B), sugars, amino acids, fatty acids, furfurals, purine alkaloids, and caffeine enable the botanical differentiation of green coffees (*4*−*9*). Caffeine, theobromine, and theophylline determined by near-infrared spectroscopy (NIRS) or liquid chromatography (LC) coupled to mass spectrometry (MS) (*10*); tocopherols and triglycerides analyzed by LC (*11*); amino acids (*6*); fatty acids (*7*) and sterols (*12*) by gas chromatography (GC) were reported for the authentication of different coffee varieties. Other approaches based on proteomics (*13*), NIRS (*14*), and Raman spectroscopy (*15*) have been also developed for this purpose. In contrast to the large amount of scientific literature dealing with the botanical origin of green coffee, the geographical characterization of coffee accounts for a reduced number of attempts. Among the most promising approaches, chlorogenic acid and cinnamoyl−amino acid conjugate contents were found to be indicators of certain geographical origins of Robusta coffees (*16*, *17*). Fatty acids, chlorogenic acids, and certain elements could distinguish between some growing areas of Arabicas (*5*); stable isotope ratio analysis of bulk green coffee (*18*) and of caffeine (*19*) enabled the determination of the geographical origin of coffees independently of their botanical origin.

The main class of phenolic compounds present in green coffee beans are chlorogenic acids (CGA), which are esters of *trans* cinnamic acids and quinic acid. Thirteen classes of CGA have been distinguished in green coffee beans (*20*−*24*): caffeoylquinic acids (CQA), feruloylquinic acids (FQA), *p*-coumaroylquinic acids (*p*CoQA), dimethoxycinnamoylquinic acid (DQA), dicaffeoylquinic acids (diCQA), diferuloylquinic acid (diFQA), di-*p*-coumaroylquinic acids, feruloylcaffeoyl quinic acids (FCQA), dimethoxycinnamoylcaffeoylquinic acid (DCQA), dimethoxycinnamoylferuloylquinic acid (DFQA), *p*-coumaroylcaffeoylquinic acids (*p*CoCQA), *p*-coumaroylferuloylquinic acids, and *p*-coumaroyldimethoxycinnamoylquinic acids. Several isomers were found in coffee due to esterification occurring at positions 3, 4, and 5, but not at position 1 (*25*). Free phenolic acids such as caffeic acid, ferulic acid, and dimethoxycinnamic acids have also been detected in green coffee extracts (*24*). These hydroxycinnamic acids also appeared in green coffee conjugated with amino acids (cinnamoyl amides) or glycosides (cinnamoyl glycosides) (*17*, *24*). All of these cinnamoyl derivatives play an important role in coffee quality, being responsible for its organoleptic properties (*26*, *27*). For instance, the quality of the beverage increases as the CGA content decreases; this fact largely explains the taste differences between Robusta and Arabica (*28*). Among methylxanthines, caffeine, theobromine, and theophylline have also been found in coffee (*10*, *24*). Caffeine is a major alkaloid in green coffee beans, the contents of which are closely related to the quality of coffee beverages, because it contributes to its bitterness (*8*, *26*).

The phenolic and methylxanthine profiles of green coffee beans are affected by several factors: coffee variety, genetic properties of the cultivars, maturity of the beans at harvest, harvesting method and postharvest processing conditions (fermentation, washing, drying, storage), agricultural practices (shade, pruning, fertilization), environmental factors (soil, altitude, sun exposure), and climatic parameters (rainfall, temperature) (*14*, *26*, *29*−*31*). Therefore, because these factors may differ from one region to the others, these chemical descriptors (concentrations of phenolic compounds and methylxanthines) are considered to be reliable geographical indicators, as well as chemotaxonomical markers. In this paper, the contents of chlorogenic acids, cinnamoyl amides, cinnamoyl glycosides, free phenolic acids and methylxanthines in green coffee beans of Arabica and Robusta varieties from the major coffee-growing regions in America, Africa, Asia, and Oceania are determined by LC coupled with UV spectrophotometry. Subsequently, the chemical data are analyzed by statistics and multivariate data analysis. The aim of this work is to present the potential of using the profiles of cinnamoyl derivatives and methylxanthines of green coffee beans together with pattern recognition techniques to distinguish coffees according to their botanical and geographical origins.

MATERIALS AND METHODS

Chemicals. Methanol (Carlo Erba, Milano, Italy) was of HPLC grade. Water was purified in a Milli-Q system from Millipore (Bedford, MA). Glacial acetic acid provided by Carlo Erba (Milano, Italy) and ascorbic acid provided by Merck (Darmstadt, Germany) were of analytical quality. All solvents used were previously filtered through 0.2 μm nylon membranes (Lida, Kenosha, WI).

Standards were supplied as follows: *p*-coumaric acid and caffeic acid by Sigma-Aldrich Chemie (Steinheim, Germany); and caffeine by Merck. Stock standard solutions at a concentration of 1 mg mL^{-1} were prepared in methanol and stored at 4 °C in darkness.

Plant Material. Green coffee beans of the *C. arabica* (Arabica) and *C. canephora* (Robusta) genera were collected from roasters and coffee dealers who were able to supply samples with an indication of the geographical origin of the coffee, at least at the national level, and from several harvests (1998−2002). The 107 green coffee samples came from different countries representing the four coffee-growing continents, that is, America, Africa, Asia, and Oceania (**Table 1**). Green coffee beans were ground and subsequently freeze-dried and stored at room temperature in a dry chamber until analysis.

Direct Solvent Extraction and Reversed-Phase HPLC Analysis. Freeze-dried coffee beans (0.1 g) were submitted to direct solvent extraction with 10 mL of methanol/water/acetic acid (30:67.5:2.5, v/v/v) with ascorbic acid (2 g/L) in an ultrasonic bath for 15 min. Then, the crude solvent extract was filtered through a 0.45 μm PTFE filter (Waters, Milford, MA) prior to injection into the HPLC system.

Chromatographic analysis was performed on a Hewlett-Packard series 1100 system, equipped with a vacuum degasser, a binary pump, a thermostated autosampler, a thermostated

Table 1. Origin of Green Coffee Bean Samples[a]

America			Africa			Asia		
country	Ara	Rob	country	Ara	Rob	country	Ara	Rob
Brazil	6		Cameroon		10	India	2	1
Colombia	1		Congo		1	Indonesia		16
Costa Rica	6		Ethiopia	7		Java		3
El Salvador	1		Kenya	1		Papua New Guinea	2	
Guatemala	9	1	Rwanda	1		Timor	1	
Honduras	1		Uganda		6	Vietnam		19
Mexico	1		Zimbabwe	1				
Nicaragua	8							
Panama	1							
Venezuela	1							

[a] Abbreviations: Ara, Arabica; Rob, Robusta.

column compartment, a photodiode array detector (DAD), and HP ChemStation software. A reversed phase Symmetry C18 (250 × 4.6 mm i.d., 5 μm) column and a Symmetry C18 (10 × 3.9 mm i.d., 5 μm) guard column (Waters) were used. The mobile phase consisted of 0.2% acetic acid in water (v/v) (solvent A) and methanol (solvent B). The elution conditions applied were the following: 0−30 min, linear gradient from 10 to 30% B; 30−40 min, linear gradient from 30 to 40% B; 40−45 min, 40% B isocratic; 45−50 min, linear gradient from 40 to 50% B; 50−55 min, 50% B isocratic; 55−65 min, linear gradient from 50 to 70% B; and finally, washing and reconditioning of the column. The flow rate was 1 mL min^{-1}, and the injection volume was 50 μL. The system operated at 25 °C. Methylxanthines were monitored and quantified at 280 nm and phenolic compounds at 320 nm. Quantification was performed by reporting the measured integration areas in the calibration equation of the standard, which exhibits a similar pattern of UV spectra (*24*). Thus, chromatographic peaks that present UV spectra with a maximum at 324−328 nm and a shoulder at 295−305 nm (compounds **2**, **3**, **5**−**9**, **11**−**15**, **17**, **18**, **20**−**25**, **27**, **28**) were reported to caffeic acid, as well as caffeoyl and feruloyl conjugates with tryptophan (two UV maxima at 290−291 and 322−323 nm); peaks with UV maximum at 311−317 nm (compounds **10**, **16**, **19**) and *p*-coumaroyl tryptophan (**26**) (two UV maxima at 290 and 309 nm) to *p*-coumaric acid; and methylxanthines (**1**, **4**) (UV maximum at 271−273 nm) to caffeine.

Identification of the compounds present in the chromatographic peaks is described elsewhere (*24*). Peak assignments were performed on the basis of the UV spectrum, retention time, and mass spectra (MS^1 and MS^2 in positive and negative ion modes) obtained by liquid chromatography coupled with a photodiode array detector, electrospray ionization, collision-induced dissociation, and tandem mass spectrometry and by comparison with the commercial standards available and/or bibliographic sources.

Data Analysis and Chemometric Procedures. Each data set consisted of a matrix in which rows represented the green coffee bean samples (objects) and columns the concentration of individual phenolic compounds and methylxanthines determined by HPLC-DAD (variables). Each sample was represented in the *n*-dimensional space by a data vector, which is an assembly of the *n* features. Data vectors belonging to the same class or category (botanical or geographical origins) were analyzed using univariate procedures [two way-ANOVA (analysis of variance), Fisher index ,and box−whisker plots] and, if necessary, chemometric techniques: unsupervised as principal component analysis (PCA) and supervised as linear discriminant analysis (LDA), partial least-squares discriminant analysis (PLS-DA), and classification trees (CART). These multivariate techniques are reviewed by Berrueta et al. (*32*). Statistic and chemometric data analyses were performed by statistical software packages Statistica 6.1 (StatSoft Inc., Tulsa, OK, 1984−2004), The Unscrambler 9.1 (Camo Process AS, Oslo, Norway, 1986−2004), and SPSS 11.5 (SPSS for Windows, SPSS Inc., 1989−1999).

The supervised techniques were applied to the autoscaled data matrix of the phenolic and methylxanthine contents of green coffee beans. The classification rules achieved were validated by means of 3-fold cross-validation or leave-one-out cross-validation (LOO). The reliability of the classification models achieved was studied in terms of recognition ability (percentage of the members of the training set correctly classified) and prediction ability (percentage of the members of the test set correctly classified by using the models developed in the training step). The models made for each category were also evaluated in terms of sensitivity of classification (percentage of objects belonging to the category that are correctly identified by the mathematical model) and specificity of classification (the percentage of objects foreign to the category htat are classified as foreign). LDA requires performing variable selection to avoid overfitting of the classification model. Thus, a best-subset selection procedure, consisting of the selection of the five most significant variables, was carried out with each of the training-test sets of cross-validation. Then, a refined selection of the variables chosen in the previous step was performed by a stepwise-forward selection procedure.

RESULTS AND DISCUSSION

Phenolic Compounds and Methylxanthines of *C. arabica* and *C. canephora*. In this paper, IUPAC nomenclature and recommended numbering systems are used for chlorogenic acids (*33*), and common names are used for free phenolic acids, cinnamoyl amides, cinnamoyl glycosides, and methylxanthines. The profiles of phenolic compounds and methylxanthines of the green beans of the coffee varieties *C. arabica* (Arabica) and *C. canephora* (Robusta) have been previously studied (*24*). Three methylxanthines, namely, caffeine, theophylline, and theobromine, and several classes of phenolic compounds were detected in the green coffee bean extracts. The major phenolic compounds were CGA, and among them, the following were identified in the samples: three caffeoylquinic acids CQA, three FQA, one *p*CoQA, three diCQA, three FCQA, four *p*CoCQA, three diFQA, six DCQA, and six DFQA (*24*). Moreover, three *trans*-cinnamic acids (caffeic acid, ferulic acid, and dimethoxycinnamic acid), six cinnamoyl−amino acid conjugates (caffeoyl-*N*-tyrosine, *p*-coumaroyl-*N*-tyrosine, caffeoyl-*N*-tryptophan, *p*-coumaroyl-*N*-tryptophan, feruloyl-*N*-tryptophan, caffeoyl-*N*-phenylalanine), and three cinnamoyl glycosides (caffeoylhexose, dicaffeoylhexose, and dimethoxycinnamoylhexose) were also determined in the extracts (*24*).

The contents of chlorogenic acids, cinnamoyl amides, cinnamoyl glycosides, free phenolic acids, and methylxanthines of the green coffee bean extracts determined by HPLC-DAD and used for the botanical characterization of coffee are summarized in **Table 2**. Compounds overlapping in the same chromatographic peak, as occurred in peaks 22, 23, and 27, were quantified together. Some compounds were present at trace levels, below the limit of quantification, or overlapped with unknown substances, which made their quantification unfeasible; therefore, these compounds were neither quantified nor considered in the data analysis performed for classification purposes.

The composition of green coffee beans varies greatly between species (*16*, *27*), as is also observed in **Table 2**. Robusta green coffee beans presented the highest total content of phenolic compounds and methylxanthines. Robustas contained higher amounts of the two methylxanthines quantified, as well as the majority of the phenolic compounds, as observed in previous studies performed on green coffee beans (*4*, *28*, *30*, *34*). This fact makes Arabica more vulnerable to phytopathogens as well as to biological and mechanical stress than Robusta (*35*). The cinnamoyl derivatives 5-*p*CoQA (**10**), 3-*p*Co,5-CQA (**19**), and dimethoxycinnamoylhexose (**25**) were the only compounds present in similar or higher concentrations in Arabicas. Caffeine (**4**) and theophylline (**1**) were also found in green coffee beans in quantities similar to those in previous studies (*8*, *30*), as well as the contents of CQA, FQA, di-CQA, and FCQA (*28*, *36*, *37*). The known large variability of Robustas from a genetic point of view is also reflected in their composition (*2*). Indeed, the concentration ranges found in Robusta were considerably larger than those in Arabica. Thus, Arabicas presented a more homogeneous composition independent of their geographical origin, explained by the low genetic diversity characteristic of this species.

Table 2. Concentrations of Phenolic Compounds and Methylxanthines (Milligrams per Kilogram of DW) in Arabica and Robusta Green Coffee Beans[a]

		Arabica (*n* = 50)				Robusta (*n* = 57)			
peak	compound	mean	SD	min	max	mean	SD	min	max
1	theophylline	nd				89	126	nd	832
2	3-CQA	2448	624	1464	3983	3938	688	2656	5686
3	dicaffeoylhexose	13	5	7	31	35	26	12	178
4	caffeine	15075	1630	10533	17412	26684	2401	20224	31582
5	5-CQA	31921	4704	22270	45325	32786	4992	21621	42903
6	caffeic acid	nd				647	236	333	1414
7	caffeoylhexose	nd				21	15	nd	47
8	3-FQA	234	54	105	358	698	154	433	1105
9	4-CQA	3328	615	2106	4825	4919	709	3670	6806
10	5-*p*CoQA	158	49	72	268	67	19	25	116
11	5-FQA	2005	276	1493	2631	5950	742	4147	8271
12	ferulic acid	43	18	22	103	221	61	61	357
13	4-FQA	229	66	79	370	840	177	581	1352
14	caffeoyl-*N*-tyrosine	nd				56	165	nd	602
15	3,5-diCQA	2559	658	1463	3892	2901	546	2018	4410
16	*p*-coumaroyl-*N*-tyrosine	nd				18	58	nd	254
17	dimethoxycinnamic acid	nd				6	31	nd	205
18	3,4-diCQA	1009	265	630	1682	3340	387	2443	4074
19	3-*p*Co,5-CQA	60	21	23	121	44	14	9	74
20	3-C,5-FQA	155	56	60	276	596	79	402	752
21	4,5-diCQA	1069	277	614	1767	2878	477	1628	3756
22	3-C,4-FQA and 3-*p*Co,4-CQA	nd				685	76	513	868
23	caffeoyl-*N*-tryptophan and 3-C,5-DQA	170	65	67	340	1557	258	1104	2389
24	caffeoyl-*N*-phenylalanine	nd				26	77	nd	274
25	dimethoxycinnamoyl-hexose	31	13	nd	66	nd			
26	*p*-coumaroyl-*N*-tryptophan	10	7	3	36	225	44	119	352
27	feruloyl-*N*-tryptophan and 3-C,4-DQA	nd				39	15	17	92
28	3-D,5-FQA	nd				7	20	nd	73

[a] Abbreviations: DW, dry weight; max, maximum; min, minimum; nd, not detected; SD, standard deviation; 3-CQA, 3-caffeoylquinic acid; 4-CQA, 4-caffeoylquinic acid; 5-CQA, 5-caffeoylquinic acid; 3-FQA, 3-feruloylquinic acid; 4-FQA, 4-feruloylquinic acid; 5-FQA, 5-feruloylquinic acid; 5-*p*CoQA, 5-*p*-coumaroylquinic acid; 3,5-diCQA, 3,5-dicaffeoylquinic acid; 3,4-diCQA, 3,4-dicaffeoylquinic acid; 4,5-diCQA, 4,5-dicaffeoylquinic acid; 3-*p*Co,5-CQA, 3-*p*-coumaroyl-5-caffeoylquinic acid; 3-*p*Co,4-CQA, 3-*p*-coumaroyl-4-caffeoylquinic acid; 3-C,5-FQA, 3-caffeoyl-5-feruloylquinic acid; 3-C,4-FQA, 3-caffeoyl-4-feruloylquinic acid; 3-D,5-FQA, 3-dimethoxycinnamoyl-5-feruloylquinic acid; 3-C,5-DQA, 3-caffeoyl-5-dimethoxycinnamoylquinic acid; 3-C,4-DQA, 3-caffeoyl-4-dimethoxycinnamoylquinic acid.

5-Caffeoylquinic acid (5-CQA) (**5**) was the major phenolic compound in green coffee beans, being present in similar amounts in both coffee varieties [21−45 g/kg of dried weight (DW) green beans]. However, the contents of several other compounds were very dependent on the coffee variety. On the one hand, caffeic acid (**6**), 3-C,4-FQA (the major component in the peak) and 3-*p*Co,4-CQA (**22**), and feruloyl-*N*-tryptophan (the major component in the peak) and 3-C,4-DQA (**27**) were detected in all Robusta samples, but not in Arabica. Moreover, caffeine (**4**), 3-FQA (**8**), 5-FQA (**11**), 4-FQA (**13**), 3,4-diCQA (**18**), 3-C,5-FQA (**20**), caffeoyl-*N*-tryptophan (the major component in the peak) and 3-C,5-DQA (**23**), and *p*-coumaroyl-*N*-tryptophan (**26**) were present at higher concentration levels in Robusta than in Arabica, allowing the unequivocal botanical distinction of green coffee beans of both varieties. Cinnamoyl glycoside (**3**, **7**, **25**) contents in green coffee beans are reported here for the first time: 7−97 mg/kg of DW of Arabica green coffee beans and 12−226 mg/kg of DW of Robusta green beans. Dimethoxycinnamoylhexose (**25**) was detected in only Arabica samples, whereas theophylline (**1**) and caffeoylhexose (**7**) were detected only in Robusta. Clifford et al. (*30*) also observed theophylline only in the Robusta species. To the authors' knowledge, the contents of dimethoxycinnamic acid (**17**) and its diacylquinic esters (**23**, **27**, **28**) in green coffee beans are reported here for the first time.

The chemical composition of coffee beans is influenced not only by genetic but also by technological and environmental factors, which are indirectly related to the geographical area of production. The contents of CQA, FQA, and di-CQA determined in African Robusta (**Table 3**) were within the concentration ranges found in the bibliography (*8*) for this coffee variety cultivated in Africa. However, the highest levels of phenolic compounds reported by Ky et al. (*8*) were not found in any of the samples studied. This observation can be explained by the fact that some of the geographical origins of the samples studied in the present work and in the study performed by Ky et al. (*8*) were different. The amounts of cinnamoyl−amino acid conjugates in the beans have also been reported to depend on the geographical origin of the coffee (*17*, *34*). The contents of such compounds determined in Robusta green beans from Angola (*16*) and India (*38*) were 3−4 times higher than the quantities found in the green coffee beans analyzed in the present study. This can be justified by the fact that the samples were from different origins (in the present study, only one coffee was from India and there were none from Angola). African Arabica contents were comparable with those reported before for Arabica green coffee beans cultivated in Ethiopia and Kenya (*8*); in this case, the same coffee-growing areas were considered in both studies.

Robusta green coffee beans from certain origins contain characteristic phenolic profiles as observed here for Ugandan green coffee beans and as had been reported before for Angolan coffee (*34*). *p*-Coumaroyl-*N*-tyrosine (**16**), caffeoyl-*N*-phenylalanine (**24**), dimethoxycinnamic acid (**17**), and 3-D,5-FQA (**28**) were detected only in Robustas from Uganda. Caffeoyl-*N*-tyrosine (**14**) was present in Ugandan green

Table 3. Concentrations of Phenolic Compounds and Methylxanthines (Milligrams per Kilogram of DW) in Robusta Green Coffee Beans from Africa (AF) and Asia−Oceania (AO)[a]

		AF (*n* = 17)				AO (*n* = 39)			
peak	compound	mean	SD	min	max	mean	SD	min	max
1	theophylline	120	200	nd	832	77	75	nd	288
2	3-CQA	4044	551	3216	5557	3905	747	2656	5686
3	dicaffeoylhexose	42	36	17	178	33	20	13	98
4	caffeine	25359	2433	20224	28436	27174	2158	23923	31582
5	5-CQA	29482	4619	21621	36255	33996	4385	25015	42903
6	caffeic acid	796	295	378	1414	589	172	333	928
7	caffeoylhexose	29	14	nd	47	17	14	nd	47
8	3-FQA	702	135	526	1105	699	164	433	1045
9	4-CQA	4957	686	3866	6806	4913	732	3670	6638
10	5-*p*CoQA	61	24	25	116	70	16	44	110
11	5-FQA	5320	626	4147	6162	6206	622	5015	8271
12	ferulic acid	254	65	100	357	208	54	61	313
13	4-FQA	816	175	603	1352	853	180	581	1257
15	3,5-diCQA	2577	398	2043	3548	3045	552	2018	4410
18	3,4-diCQA	3191	450	2443	4074	3409	346	2649	4039
19	3-*p*Co,5-CQA	36	10	15	49	48	13	9	74
20	3-C,5-FQA	539	66	402	665	623	71	495	752
21	4,5-diCQA	2845	521	1628	3756	2907	458	2017	3714
22	3-C,4-FQA and 3-*p*Co,4-CQA	657	75	513	768	698	75	562	868
23	caffeoyl-*N*-tryptophan and 3-C,5-DQA	1665	283	1198	2099	1513	239	1104	2389
26	*p*-coumaroyl-*N*-tryptophan	207	39	119	276	231	44	152	352
27	feruloyl-*N*-tryptophan and 3-C,4-DQA	43	19	17	79	38	14	17	92

[a] Abbreviations: see **Table 2**.

coffee beans (300−602 mg/kg of DW) and in the only sample from Congo (67 mg/kg of DW). *p*-Coumaroyl-*N*-tyrosine (**16**) had been previously detected only in Angolan Robustas [1.3−4.8 g/kg of DW (*17*)], which presented concentrations 10 times higher than those found in Ugandan samples (0.1−0.3 g/kg of DW) in the present study. As a result, *p*-coumaroyl-*N*-tyrosine is considered to be a potential reliable indicator of the origin of Robustas from Angola and Uganda.

Statistical Data Analysis. *Botanical Classification: C. arabica and C. canephora.* The data set was made up of 107 samples of green coffee beans and 28 variables, which were the concentrations of phenolic compounds and methylxanthines determined by HPLC-DAD. The analysis of variance (ANOVA) performed on this matrix disclosed that the contents of caffeine and some phenolic compounds of the green beans of both coffee varieties, Arabica and Robusta, were significantly different. The Fisher index was calculated to establish the discriminant capacity of the variables one by one (*39*). The variables that presented the highest Fisher weights ($p < 0.02$) and for which box and whisker plots showed complete separation of the concentration ranges in the two coffee varieties were totally discriminant and, therefore, allowed the botanical characterization of the green coffee beans. These variables were *p*-coumaroyl-*N*-tryptophan (**26**), caffeoyl-*N*-tryptophan and 3-C,5-DQA (**23**), 3-FQA (**8**), 5-FQA (**11**), 4-FQA (**13**), caffeine (**4**), 3,4-diCQA (**18**), and 3-C,5-FQA (**20**) (**Table 2**). Moreover, caffeic acid (**6**), 3-C,4-FQA and 3-*p*Co,4-CQA (**22**), and feruloyl-*N*-tryptophan and 3-C,4-DQA (**27**) were detected in all of the Robusta samples but not in any of the Arabica samples (**Table 2**), so they were also completely discriminant for the botanical differentiation of both coffee varieties. In contrast, the total phenolic contents of the green coffee beans were not completely discriminant between the two coffee varieties.

The variations in the phenolic composition of coffee varieties are mainly due to genetic factors, but these compounds are also sensitive to all types of environmental changes and, consequently, interactions should exist when genotypes are grown in different regions (*27*). All of these genetic and environmental factors affect the biosynthetic pathways of CGA formation (*27*, *36*), so different chemical and enzymatic reactions can take place in the coffee beans, resulting in the aforementioned variabilities in their chemical composition. In this context and with regard to the contents of the compounds that discriminate between the two coffee species studied, it is concluded that the biosynthesis of FQA is favored in Robustas, which is reflected in the high concentrations of the three monoesters of FQA (**8**, **11**, **13**) and the diesters 3-C,5FQA (**20**) and 3-C,4-FQA (**22**). Besides this, the conjugation of *trans*-cinnamic acids with the amino acid tryptophan (**23**, **26**, **27**), the cinnamoyl substitution reactions in the 3-position, and the subsequent transesterifications (**18**, **20**, **22**), as well as caffeine (**4**) biosynthesis, all occur to a greater extent in Robusta than in Arabica.

Thus, the profiles of chlorogenic acids, cinnamoyl amides, cinnamoyl glycosides, free phenolic acids, and methylxanthines have proven to be useful as chemotaxonomic tools for the botanical characterization of both *Coffea* species, that is, *C. arabica* (Arabica) and *C. canephora* (Robusta). In further studies, the botanical maturity of the green beans should be taken into consideration to evaluate its influence on the coffee bean composition (*29*, *30*).

Geographical Origin of C. arabica. (a) Continential Origin. ANOVA performed on the data set of Arabica green beans, which contained 18 variables (concentrations of some cinnamate conjugates and caffeine) and 50 samples, revealed that there were significant differences for most of the variables between the coffee-growing continents [America (AM), Africa (AF), and Asia−Oceania (AO)] (**Table 4**). A least significant difference (LSD) test ($p < 0.05$) was also carried out on the data matrix, to confirm that there were no significant differences between years of harvest. The Fisher test allowed us to detect the most discriminant

Table 4. Concentrations of Phenolic Compounds and Methylxanthines (Milligrams per Kilogram of DW) in Arabica Green Coffee Beans from America (AM), Africa (AF), and Asia−Oceania (AO)[a]

		AM (n = 35)				AF (n = 10)				AO (n = 5)			
peak	compound	mean	SD	min	max	mean	SD	min	max	mean	SD	min	max
2	3-CQA	2511	643	1464	3931	2056	273	1490	2511	2794	722	2214	3983
3	dicaffeoylhexose	13	4	8	29	14	4	9	22	14	9	7	31
4	caffeine	15239	1635	10533	17412	14162	1592	12215	16757	15752	1112	14312	16874
5	5-CQA	30412	3774	22270	38454	36024	5887	29878	45325	34277	1759	32731	36614
8	3-FQA	246	51	167	358	184	41	105	242	252	34	214	305
9	4-CQA	3361	633	2106	4797	2965	306	2549	3468	3821	614	3379	4825
10	5-*p*CoQA	163	53	72	250	137	23	105	173	162	62	104	268
11	5-FQA	1985	249	1498	2421	2010	292	1675	2406	2136	434	1493	2631
12	ferulic acid	40	15	22	93	52	16	27	76	46	32	25	103
13	4-FQA	239	64	95	370	176	53	79	234	259	49	193	307
15	3,5-diCQA	2609	623	1757	3892	2341	688	1463	3701	2637	888	1644	3475
18	3,4-diCQA	1064	272	660	1682	770	106	630	942	1107	155	961	1291
19	3-*p*Co,5-CQA	64	23	37	121	46	8	39	59	54	19	23	75
20	3-C,5-FQA	164	56	92	276	122	53	60	200	163	51	98	206
21	4,5-diCQA	1078	284	614	1767	1010	248	655	1454	1127	327	854	1641
23	caffeoyl-*N*-tryptophan and 3-C,5-DQA	162	61	67	275	159	54	82	211	245	72	166	340
25	dimethoxycinnamoylhexose	33	11	16	63	23	19	0	66	33	12	15	45
26	*p*-coumaroyl-*N*-tryptophan	9	4	3	21	9	4	4	16	20	14	7	36

[a] Abbreviations: see **Table 2**.

variables ($p < 0.01$) between continents (*39*), which were 3-*p*Co,5-CQA (**19**) and 3,4-diCQA (**18**) between AF and AM, *p*-coumaroyl-*N*-tryptophan (**26**) between AF and AO, and *p*-coumaroyl-*N*-tryptophan (**26**) and dicaffeoylhexose (**3**) between AM and AO. The discriminant capacity of these variables was due to their variability in each category. In this sense, the variability of the amounts of compounds **26** and **3** in AO green coffee beans was larger than in AF and AM; the variability of compounds **18** and **19** was more prominent in AM than in AF (**Table 4**). However, the box−whisker plots of these variables showed an overlap in the concentration ranges of these compounds; thus, none of the variables measured was able, by itself, to discriminate the Arabica samples from the three origins. For this reason, it was necessary to apply multivariate data analysis to achieve the desired differentiation.

PCA was performed on the Arabica data set. The three first principal components accounted for 72% of total system variability. The bidimensional plots of the sample scores in the space defined by the first principal component (PC1, 32% of total variability) versus the second principal component (PC2, 26% of total variability) indicated a natural separation of AF and AO green coffee beans (**Figure 1a**). PC1 was responsible for the distinction between AF and AO samples. The loadings of the variables showed that 4-CQA (**9**), 3,4-diCQA (**18**), 4-FQA (**13**), 3-FQA (**8**), and 3-CQA (**2**) were the most influential features on PC1, due to the higher contents of these phenolic compounds in AO samples than in the AF ones. Thus, the reaction of transesterification of CGAs was enhanced by AO environmental, climatic, and agricultural factors, which were probably the factors mainly responsible for the difference between the Arabicas from these two geographical origins. In contrast, all Arabica AF and AO samples completely overlapped with the AM category. LDA and PLS-DA were applied to the data of the present sample set to produce classification models for the geographical characterization of the Arabica green coffee beans as a preliminary attempt. The classification results achieved are shown in **Table 5**. The 3-fold cross-validated LDA model correctly classified 94.3% of AM Arabica samples, whereas the recognition and prediction abilities of the model for AF and AO samples were quite poor, less than 65% of hits. Besides the low percentages of correct classifications, the recognition ability was lower than the prediction ability when two of the three CV sets were used. Therefore, classification results strongly depend on the samples in the training and test sets used in the cross-validation, which was due to the unbalanced number of samples in each class and the low number of representatives in the AO and AF categories. As a consequence, the LDA model was not stable and, therefore, unreliable.

Instead of *k*-fold cross-validation, LOO cross-validation is recommended for small database sizes that present the problem of the inability to divide the data set into fairly sized subsets for training and test sets (*32*). To improve the final model, PLS-DA models were validated by LOO. Better results were obtained for AM and AF Arabica green coffee beans, but not for AO, as expected, because of the low number of samples. Two variants of PLS-DA were carried out: (*i*) a model is computed for each class separately so that it distinguishes its samples from all others (class-modeling PLS, PLS-*i*); and (*ii*) all classes were modeled simultaneously (classical PLS, PLS-*ii*) (*40*). The same classification abilities were obtained with the two approaches for AF samples (recognition and prediction abilities: 90 and 70%, respectively). For AM, similar results were achieved for both procedures: in PLS-*i*, recognition and prediction abilities were 97 and 86%, respectively; and in PLS-*ii*, 97% for both. Neither PLS approach manages to provide a model for category AO, demonstrating that there was not enough information of AO Arabicas for the sought differentiation. The specificity of the AF (PLS-*i*) model was 97% to AM green coffee beans, so only 3% of AM samples were wrongly classified by the model as AF. This model could be used in fraud detection (AM coffees passed as AF). Furthermore, the AM model could also be used for this purpose, even though this model misclassified 10% of the AF coffees as authentic AM.

From the weighted regression coefficients, the most influential variables in the PLS-DA models were **5** (5-CQA), **25**

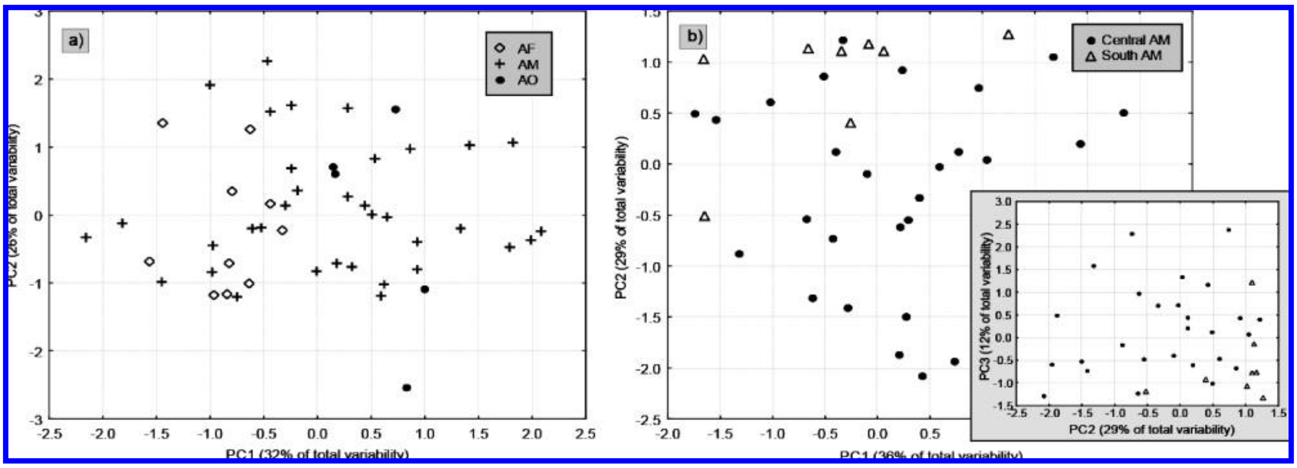

Figure 1. Projection on the multidimensional space defined by the principal components: (**a**) Arabica coffees, identifying the samples according to their origin at the continent level (AM, AF, AO); (**b**) American Arabica coffees, identifying the samples according to their region of origin (Central and South America).

Table 5. Classification Results for the Supervised Pattern Recognition Techniques Applied to the Data of Phenolic and Methylxanthine Contents of Arabica Green Coffee Beans for Their Distinction at the Continental Level[a]

					PLS-DA[b]				
			LDA[c]		PLS-*ii*[d]		PLS-*i*[e]		
origin	*N*	a priori prob	recog (%)	predic (%)	recog (%)	predic (%)	recog (%)	predic (%)	specificity to (%)
AF	10	0.21	65	60	90	70	90	70	AM: 97 AO: 60
AM	35	0.70	94	94	97	97	97	86	AF: 90 AO: 60
AO	5	0.09	40	40	40	20			

[a] Abbreviations: *N*, number of samples; prob, probability; recog, recognition ability; predic, prediction ability. [b] PLS-DA performed on The Unscrambler; LOO validation. [c] LDA performed on Statistica using four compounds (**5**, **18**, **25**, **26**), 3-fold cross-validation. [d] PLS-*ii*, classical PLS. Sample codification: AF (1,0,0), AM (0,1,0), AO (0,0,1); 3 PLS components selected; borders at 0.3000, 0.6750, and 0.1275, respectively. [e] PLS-*i* class-modeling. PLS: 3 PLS components selected; border of AF model at 0.3000; border of AM model at 0.6300.

(dimethoxycinnamoylhexose), **8** (3-FQA), and **18** (3,4-diCQA) in AF (PLS-*i* and PLS-*ii* models) and **5** (5-CQA) and **4** (caffeine) in AM (PLS-*i* model). The optimized and validated LDA model also included some of these variables (**5**, **18**, **25**), together with **26** (*p*-coumaroyl-*N*-tryptophan). AF Arabicas presented relatively higher contents of 5-CQA and lower 3,4-diCQA compared to the AM ones; that is, in AM Arabica the biosynthetic pathway of diCQA is favored. Moreover, the lower contents of 3-FQA in AF Arabicas also indicate that the reactions involved in the FQA formation pathway take place at different rates depending on the geographical origin and, because 3-FQA biosynthesis is genetically controlled (*27*), that the Arabica genotypes grown in AF may be different. AO Arabicas were characterized by high levels of *p*-coumaroyl-*N*-tryptophan. The fact that models achieved by different techniques and approaches were based on the same variables implies that the results were feasible and not random. The multivariate data analysis results disclosed that phenolic and methylxanthine profiles of green coffee beans may contain adequate information to achieve the differentiation of Arabica coffee according to its geographical origin. However, to develop suitable tools for the geographical characterization of these coffees, further studies should be performed with sample sets containing more samples and a balanced number of representatives in each category.

(b) National Origin. With regard to coffee characterization at the country level, univariate data analysis of the Arabica data set showed that the phenolic compounds studied and caffeine did not completely discriminate between Arabica green coffee beans from the different countries, even though the single sample of some countries exhibited significantly different amounts of one or more compounds with respect to the samples of the other countries (data not shown). This was the case for the sample from Timor, which presented the highest amounts of **3** (dicaffeoylhexose), **12** (ferulic acid), **23** (caffeoyl-*N*-tryptophan and 3-C,5-DQA), and **26** (*p*-coumaroyl-*N*-tryptophan) and the lowest amount of **19** (3-*p*Co,5-CQA). Analogously, the samples from Venezuela and El Salvador contained the lowest concentrations of **4** (caffeine), **5** (5-CQA), and the total phenolic compounds and methylxanthines, and that from El Salvador showed also the lowest amount of **9** (4-CQA). The samples from Colombia and Kenya showed the highest contents of **21** (4,5-diCQA) and **25** (dimethoxycinnamoylhexose), respectively. These results were preliminary because only one sample from each of these origins was available for this study. Therefore, no statements can be made regarding this issue, and further studies should be performed with a representative number of samples for each country.

With regard to the aforementioned importance of American Arabica coffees, the AM Arabica data set (**Table 6**) was further analyzed by PCA. The score plot defined by the first two principal components (accounting for 65% of total variability of the system) disclosed some correlation with the geographical origin of the coffee-growing countries: Central America (Panama, Costa Rica, Nicaragua, Honduras, El Salvador,

Table 6. Concentrations of Phenolic Compounds and Methylxanthines (Milligrams per Kilogram of DW) in Arabica Green Coffee Beans from Central and South America[a]

		Central America (n = 27)				South America (n = 8)			
peak	compound	mean	SD	min	max	mean	SD	min	max
2	3-CQA	2367	627	1464	3931	2997	445	2157	3777
3	dicaffeoylhexose	13	5	8	29	13	2	11	15
4	caffeine	15332	1546	10898	17412	14924	1992	10533	17203
5	5-CQA	30768	3863	22270	38454	29211	3407	22803	33150
8	3-FQA	231	46	167	338	295	37	232	358
9	4-CQA	3250	635	2106	4797	3734	494	2751	4518
10	5-*p*CoQA	160	53	72	250	175	55	111	248
11	5-FQA	1982	262	1498	2421	1994	213	1689	2356
12	ferulic acid	39	16	22	93	44	12	31	69
13	4-FQA	224	62	95	332	289	49	225	370
15	3,5-diCQA	2760	592	1777	3892	2102	446	1757	3160
18	3,4-diCQA	1023	287	660	1682	1202	157	982	1466
19	3-*p*Co,5-CQA	64	21	37	111	67	29	40	121
20	3-C,5-FQA	168	56	92	276	149	57	100	254
21	4,5-diCQA	1047	276	614	1634	1181	305	823	1767
23	caffeoyl-*N*-tryptophan and 3-C,5-DQA	177	59	82	275	112	38	67	168
25	dimethoxycinnamoylhexose	37	10	23	63	22	6	16	33
26	*p*-coumaroyl-*N*-tryptophan	10	4	4	21	6	2	3	10

[a] Abbreviations: see **Table 2**.

Table 7. Classification Results for Supervised Pattern Recognition Techniques Applied to the Data of Phenolic and Methylxanthine Contents of American Arabica Green Coffee Beans for Their Distinction at the Subcontinental Level[a]

			LDA[b]		PLS-DA[c]		PLS-DA[d]	
origin	N	a priori prob	recog (%)	predic (%)	recog (%)	predic (%)	recog (%)	predic (%)
Central AM	28	0.75	98	93	94	89	96	82
South AM	7	0.25	75	75	94	88	88	75

[a] Abbreviations: see **Table 5**. [b] LDA performed on Statistica using two compounds (**23**, **25**), 3-fold cross-validation. [c] PLS-DA performed by Statistica, 3-fold cross-validation: CV-1, two PLS components selected and border at 0.4450; CV-2, one PLS-component selected and border at 0.4125; CV-3, three PLS components selected and border at 0.3900. [d] PLS-DA performed by The Unscrambler, LOO validation; two PLS components selected and border at 0.4125.

Guatemala, and Mexico) and South America (Colombia, Venezuela, and Brazil) (**Figure 1b**). Despite the unbalanced number of samples in each category, pattern recognition techniques achieved similar and interesting results (**Table 7**). However, due to these unbalances, the classifications achieved by the LDA model were biased to the class with the higher number of representatives (i.e., Central AM), and those by the PLS-DA models were dependent on the samples in the training and test sets. The data contained some substantial information related to the geographical origin of coffees, even though these results are expected to improve by using a balanced and representative data set.

Geographical Origin of C. canephora

(a) Continental Origin. The Robusta coffee data set consisted of 56 green coffee bean samples from AF and AO and only one from AM characterized by 27 variables (phenolic compound and methylxanthine concentrations). No significant differences were detected between samples from different years of harvest [LSD test (p < 0.05)]. The most discriminant variables between AF and AO Robusta samples were theophylline (**1**), dicaffeoylhexose (**3**), and caffeic acid (**6**) [Fisher test (p < 0.01)], but the box−whisker plots of these variables showed that the concentration ranges in AF and AO categories overlapped, AF presenting larger concentration ranges than AO (**Table 3**).

(b) National Origin. With regard to the Robusta samples at the national level, some compounds were detected in the coffee from only certain countries. This occurred in Robusta green beans from Uganda, which were the only samples containing *p*-coumaroyl-*N*-tyrosine (**16**), caffeoyl-*N*-phenylalanine (**24**), 3-D,5-FQA (**28**), and dimethoxycinnamic acid (**17**), whereas caffeoyl-*N*-tyrosine (**14**) was also detected in the sample from Congo but at a much lower concentration than in the Ugandan samples, as already mentioned above. *p*-Coumaroyl-*N*-tyrosine had been also found in Robusta beans from Angola by Clifford et al. (*17*), but at significantly higher concentrations than in Robusta green coffee beans from Uganda in the present work. As a result, these cinnamoyl derivatives are claimed to be reliable indicators of the geographical origin for Robustas from Angola and Uganda.

The above observations have to be taken with caution because only one sample from Congo, one from India, and three from Java were studied, and they cannot be considered as representative for these countries. The Congolese Robusta sample presented distinctively high contents of theophylline (**1**), dicaffeoylhexose (**3**), 3-FQA (**8**), 4-CQA (**9**), and 4-FQA (**13**) in comparison with the samples from the other origins. The Indian Robusta sample contained relatively high amounts of 3,5-diCQA (**15**), caffeoyl-*N*-tryptophan and 3-C,5-DQA (**23**), and feruloyl-*N*-tryptophan and 3-C,4-DQA (**27**). Robustas from Java were peculiar due to their high amounts of 5-FQA (**11**). These preliminary investigations should be completed with representative sample sets from these countries.

PCA performed on the complete data set of Robusta samples confirmed that Ugandan samples formed a separate

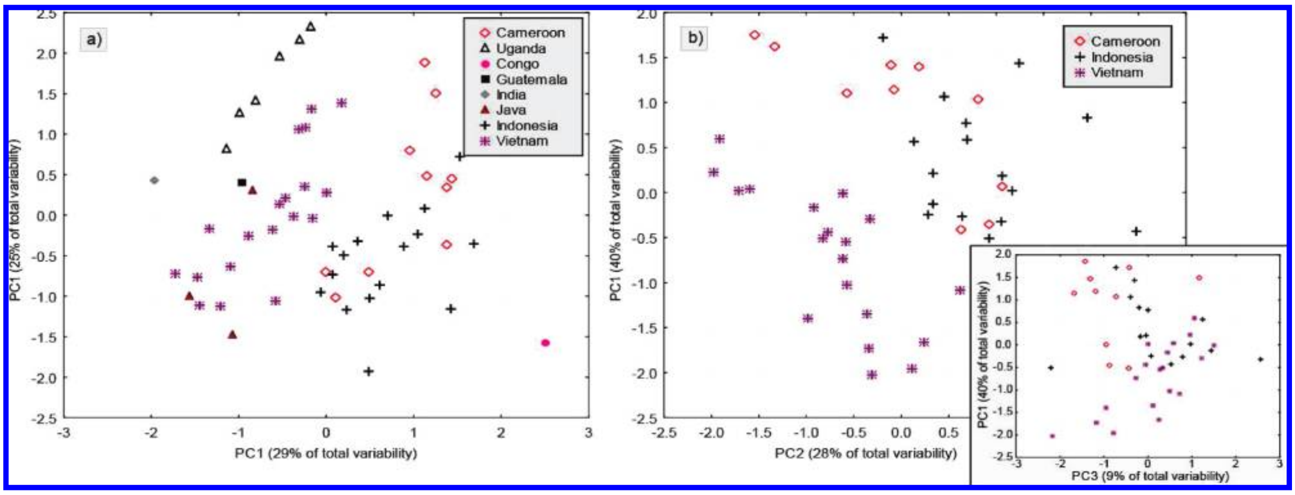

Figure 2. Projection of the Robusta coffees on the multidimensional space defined by the principal components, identifying the samples according to their country of origin: (**a**) all Robusta coffees; (**b**) coffees from Cameroon, Indonesia, and Vietnam.

Table 8. Concentrations of Phenolic Compounds and Methylxanthines (Milligrams per Kilogram of DW) in Robusta Green Coffee Beans from Cameroon, Indonesia, and Vietnam[a]

		Cameroon (n = 10)				Indonesia (n = 16)				Vietnam (n = 19)			
peak	compound	mean	SD	min	max	mean	SD	min	max	mean	SD	min	max
1	theophylline	91	74	34	285	134	87	41	288	36	19	nd	70
2	3-CQA	4022	450	3216	4660	4642	463	4057	5686	3357	369	2656	4028
3	dicaffeoylhexose	35	10	24	60	51	19	23	98	22	7	14	44
4	caffeine	24534	2533	20224	28347	25733	1364	23923	28661	27855	1979	24425	31198
5	5-CQA	27353	4197	21621	34929	30484	3030	25015	35600	35827	3262	30195	40527
6	caffeic acid	920	226	632	1414	749	145	505	928	502	56	415	605
7	caffeoylhexose	29	15	nd	47	24	8	13	35	16	17	nd	47
8	3-FQA	715	86	542	830	867	87	760	1045	573	63	468	727
9	4-CQA	4851	563	3866	5678	5592	506	4826	6638	4383	401	3670	5172
10	5-*p*CoQA	46	11	25	61	66	18	44	100	73	15	50	110
11	5-FQA	5050	609	4147	6093	5985	507	5015	6651	6180	448	5364	6956
12	ferulic acid	285	41	218	357	241	41	174	313	208	14	183	241
13	4-FQA	828	107	613	990	1028	100	893	1257	712	75	591	893
15	3,5-diCQA	2713	424	2285	3548	2715	331	2018	3309	3230	544	2565	4193
18	3,4-diCQA	3315	487	2527	4074	3352	275	2649	3765	3390	383	2661	3998
19	3-*p*Co,5-CQA	36	8	16	44	46	9	31	65	54	9	40	74
20	3-C,5-FQA	528	68	402	637	613	60	495	718	624	79	514	749
21	4,5-diCQA	3034	460	2346	3756	2933	326	2262	3484	2945	517	2052	3714
22	3-C,4-FQA and 3-*p*Co,4-CQA	646	92	513	768	706	72	582	822	679	61	562	809
23	caffeoyl-*N*-tryptophan and 3-C,5-DQA	1594	262	1217	2024	1519	173	1154	1915	1496	204	1104	2024
26	*p*-coumaroyl-*N*-tryptophan	187	35	119	238	203	31	152	284	243	37	194	352
27	feruloyl-*N*-tryptophan and 3-C,4-DQA	31	7	17	38	31	9	17	54	41	10	26	63

[a] Abbreviations: see **Table 2**.

cluster in the PC1 versus PC2 plot (54% of total system variability) (**Figure 2a**). Congolese and Indian samples were far from the rest of the samples, and those from Java overlapped with the Vietnamese samples, as well as the only Robusta sample from AM.

(c) Robustas from Cameroon, Indonesia, and Vietnam. Further data analysis was performed with a data set that excluded the samples from Uganda, Congo, India, Java, and Guatemala and included only the samples from Cameroon, Indonesia, and Vietnam (coffee-growing countries well represented in the data set). Thus, the data matrix was composed of 45 green coffee bean samples and 22 variables (phenolic and methylxanthine contents). None of these variables was able to discriminate by itself between the Robusta green coffee beans from these three origins at the same time (**Table 8**). Only the caffeic acid (**6**) content enabled distinction between Robustas from Cameroon and Vietnam, but not between these two countries and Indonesia. Therefore, multivariate data analysis was needed to differentiate the coffees of these three countries.

The first three principal components represented 40, 28, and 9% of total system variability, respectively. Vietnamese Robustas were completely separated from Indonesian and Cameroonian samples, whereas samples from these two countries partially overlapped (**Figure 2b**). From the loadings of the variables, the most influential features on PC1 were 5-CQA (**5**), 3,5-diCQA (**15**), caffeine (**4**), 5-FQA (**11**), 3-C,5-FQA (**20**), 3-*p*Co,5-CQA (**19**), *p*-coumaroyl-*N*-tryptophan (**26**), and 3,4-diCQA (**18**). The major contributions to PC2 were due to 4-CQA (**9**), 4-FQA (**13**), 3-CQA (**2**), and 3-FQA (**8**). Vietnamese green coffee beans were characterized by high contents of 5-CQA, 3,5-CQA, caffeine, and

Table 9. Classification Results for the Supervised Pattern Recognition Techniques Applied to the Data of Phenolic and Methylxanthine Contents of Robusta Green Coffee Beans from Cameroon, Indonesia, and Vietnam for Their Distinction at the National Level[a]

					PLS-DA						
			LDA[b]		PLS-*ii*		PLS-*i*[e]			CART[f]	
origin	*N*	a priori prob	recog (%)	predic (%)	recog (%)	predic (%)	recog (%)	predic (%)	specificity to (%)	recog (%)	predic (%)
Cameroon	10	0.22	100	100	$100^{c,d}$	$100^{c,d}$	100	100	Indonesia: 100 Vietnam: 100	70	40
Indonesia	16	0.36	100	94	$100^{c,d}$	$100^{c,d}$	100	94	Cameroon: 100 Vietnam: 100	88	81
Vietnam	19	0.42	100	100	$100^{c,d}$	$95^{c}/100^{d}$	100	100	Cameroon: 100 Indonesia: 100	100	100

[a] Abbreviations: see **Table 5**. [b] LDA performed on Statistica using six compounds (**4**, **8**, **11**, **12**, **18**, **19**), 3-fold cross-validation. [c] PLS-DA performed on Statistica, 3-fold crossvalidation, three PLS components selected for the three CV sets and border at 0.5. [d] PLS-DA performed on The Unscrambler, LOO validation, PLS-*ii*, classical PLS: sample codification: Cameroon (1,0,0), Indonesia (0,1,0), Vietnam (0,0,1); three PLS components selected; borders at 0.5. [e] PLS-DA performed on The Unscrambler, LOO validation, PLS-*i*, class-modeling PLS: Cameroon model, two PLS components selected; Indonesia model, three PLS components selected; Vietnam model, one PLS component selected; borders at 0.500. [f] CART performed on Statistica, 3-fold cross-validation; split selection method: discriminant-based univariate splits for ordered predictors; stopping rule: prune on misclassification error; split variables: **8** and **6**.

p-coumaroyl-*N*-tryptophan; Cameroonian by low levels of 5-FQA, 3-C,5-FQA, and 3-*p*Co,5-CQA; and Indonesian by high amounts of 4-CQA, 4-FQA, 3-CQA, and 3-FQA. In Vietnamese Robustas, the transesterification of 5-CQA or 3,5-CQA is not favored, resulting in the presence of high levels of these substrates. The low levels of 5-FQA and the 3,5-dicinnamoyl isomers in Cameroonian samples can be due to the low rate of the enzymatic reaction that leads to its formation, in comparison with the samples of the other two origins. In contrast, the transesterification reaction of 5-CQA and 5-FQA is favored in Indonesian Robustas, and this is responsible for the high amounts of 3- and 4-substituted monoesters of CQA and FQA. These unsupervised results disclosed that differences exist between Robustas grown in well-separated countries from Africa and Asia, which can be due to environmental, climatic, and agricultural factors as well as genetic factors in view of the high diversity of Robusta genotypes (*2*).

Supervised pattern recognition techniques provided satisfactory models that successfully accomplished the task of distinguishing Robusta green coffee beans from the three coffee-growing areas (Cameroon, Indonesia, and Vietnam) (**Table 9**). Thus, LDA and PLS-DA afforded 3-fold cross-validated models that correctly classified all samples, except that LDA misclassified one Indonesian sample (94% of prediction ability) and PLS-*ii*, one Vietnamese sample (95% of prediction ability). Using LOO cross-validation, the same classification results as LDA were attained by PLS-*i* models, whereas the PLS-*ii* model achieved 100% of hits for the three categories. The PLS-*i* approach provided independent models for each coffee origin with very good sensitivities and specificities of classifications. The Cameroon and Vietnam models presented sensitivities of 100%, so they recognized all of the samples belonging to their own class, and specificities to the other classes, also of 100%; that is, none of the foreign samples to the models were wrongly classified by the models as belonging to their own class. The Cameroon model recognized all of the Cameroonian samples (100% of sensitivity) and did not misclassify any of the Indonesian or Vietnamese samples as Cameroonian (100% specificities to the Indonesian and Vietnamese categories). The same argument can be made for the Vietnamese model. The Indonesia model did not consider any foreign coffee as Indonesian (100% specificity to Cameroon and Vietnam categories); however, 6% of Indonesian coffees were not recognized by its own model (sensitivity of 94%). With regard to CART, this supervised technique is very influenced by unbalances in the data set; therefore, the classification results were biased toward the category with more representatives (**Figure 3**). Thus, CART classified correctly all Vietnamese samples, recognized 88% of Indonesian samples, and predicted correctly 81% of them. However, between 30 and 60% of the Cameroonian samples were misclassified.

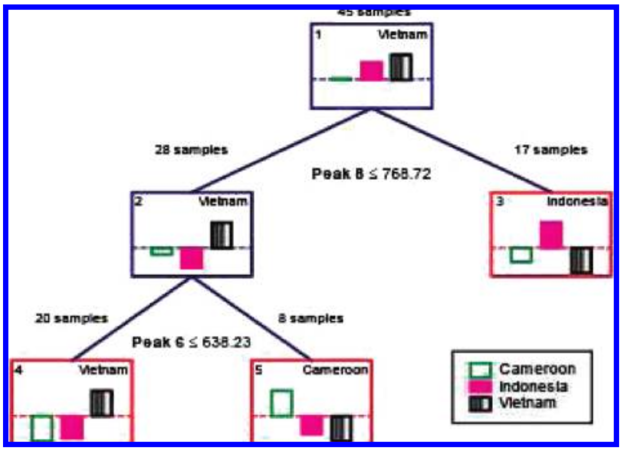

Figure 3. Classification tree for Robusta coffees from Cameroon, Indonesia, and Vietnam (two splits and three terminal nodes).

In terms of fraud detection, the LDA and PLS-DA models are very interesting because fraudulent acts (Cameroonian or Vietnamese coffees passed as Indonesian, Cameroonian or Indonesian coffees passed as Vietnamese, and Indonesian or Vietnamese coffees passed as Cameroonian) would be detected, although there would be some risk that individual Indonesian coffees would not be recognized as such. In a quality control context, the LDA and PLS-DA (PLS-*ii*) models were satisfactory because they correctly identified all authentic coffees from Cameroon and most of the samples from Indonesia or Vietnam, respectively. If more samples are included in the model-building phase, better models are expected.

The most influential features in the supervised models here developed, that is, almost all variables selected in LDA and CART and those variables with the highest weighted regression coefficients in PLS-DA, coincided with the most influential variables in PCA. This indicates that the models are reliable and stable and that the information contained in the data is adequate for the present aim, that is, the geographical characterization of Robusta green coffee beans.

The results of multivariate analysis of the green coffee data showed that the profiles of chlorogenic acids, cinnamoyl amides, cinnamoyl glycosides, free phenolic acids, and methylxanthines of green coffee beans contain adequate information for the geographical characterization of Arabica and Robusta coffees at continental, subcontinental, and national levels. Further studies are required with larger sample sets containing balanced numbers of representatives in each category, even at the *terroir* level, in order to achieve adequate tools for the geographical characterization of coffees to be used in the implementation of regulations related to denomination of origin, as well as to protect specialty coffees and promote honest competition. The detection of botanical and/or geographical adulteration in mixtures was not within the scope of the present study, but this will also be an interesting issue to be faced in future studies.

ABBREVIATIONS USED

CGA, chlorogenic acid; CQA, caffeoylquinic acid; DCQA, dimethoxycinnamoylcaffeoylquinic acid; DFQA, dimethoxycinnamoylferuloylquinic acid; diCQA, dicaffeoylquinic acid; diFQA, diferuloylquinic acid; FQA, feruloylquinic acid; FCQA, feruloylcaffeoylquinic acid; *p*CoQA, *p*-coumaroylquinic acid; *p*CoCQA, *p*-coumaroylcaffeoylquinic acid; 3-CQA, 3-caffeoylquinic acid; 4-CQA, 4-caffeoylquinic acid; 5-CQA, 5-caffeoylquinic acid; 3-FQA, 3-feruloylquinic acid; 4-FQA, 4-feruloylquinic acid; 5-FQA, 5-feruloylquinic acid; 5-*p*CoQA, 5-*p*-coumaroylquinic acid; 3,5-diCQA, 3,5-dicaffeoylquinic acid; 3,4-diCQA, 3,4-dicaffeoylquinic acid; 4,5-diCQA, 4,5-dicaffeoylquinic acid; 3-*p*Co,5-CQA, 3-*p*-coumaroyl-5-caffeoylquinic acid; 3-*p*Co,4-CQA, 3-*p*-coumaroyl-4-caffeoylquinic acid; 3-C,5-FQA, 3-caffeoyl-5-feruloylquinic acid; 3-C,4-FQA, 3-caffeoyl-4-feruloylquinic acid; 3-D,5-FQA, 3-dimethoxycinnamoyl-5-feruloylquinic acid; 3-C,5-DQA, 3-caffeoyl-5-dimethoxycinnamoylquinic acid; 3-C,4-DQA, 3-caffeoyl-4-dimethoxycinnamoylquinic acid; ANOVA, analysis of variance; LDA, linear discriminant analysis; LOO, leave one out cross-validation; LSD, least significant difference; PCA, principal component analysis; PC1, first principal component; PC2, second principal component; PC3, third principal component; PLS-DA, partial least-squares discriminant analysis; DAD, diode array detector; HPLC, high-performance liquid chromatography; DW, dry weight; SD, standard deviation; min, minimum; max, maximum; nd, not detected; Ara, Arabica; Rob, Robusta; AM, America; AF, Africa; AO, Asia-Oceania.

ACKNOWLEDGMENT

We thank Paolo Carrara of Cofitrend SRL (Genova, Italy), Gloria Pellegrino of Lavazza, Ing. Oscar García Murga (O.G. International Trading Links srl), Dr. F. Suggi Liverani, and Dr. M. Petracco (Illy S.p.a.) for the provision of coffee samples.

LITERATURE CITED

(1) Charrier, A. C.; Berthaund, J. Botanical classification of coffee. In *Coffee: Botany, Biochemistry and Production of Beans and Beverage*; Clifford, M. N., Willson, K. C., Eds.; Croom Helm: London, U.K., 1985; pp 13−47.

(2) Bertrand, B.; Guyot, B.; Anthony, F.; Lasherme, P. Impact of the *Coffea canephora* gene introgression on beverage quality of *C. arabica*. Theor. Appl. Genet. **2003**, *107*, 387–394.

(3) Mancha Agresti, P. D. C.; Franca, A. S.; Oliveira, L. S.; Augusti, R. Discrimination between defective and non-defective Brazilian coffee beans by their volatile profile. Food Chem. **2008**, *106*, 787–796.

(4) Guerrero, G.; Suarez, M.; Moreno, G. Chlorogenic acids as a potential criterion in coffee genotype selections. J. Agric. Food Chem. **2001**, *49*, 2454–2458.

(5) Bertrand, B.; Villarreal, D.; Laffargue, A.; Posada, H.; Lashermes, P.; Dussert, S. Comparison of the effectiveness of fatty acids, chlorogenic acids, and elements for the chemometric discrimination of coffee (*Coffea arabica* L.) varieties and growing origins. J. Agric. Food Chem. **2008**, *56*, 2273–2280.

(6) Casal, S.; Alves, M. R.; Mendes, E.; Oliveira, M. B. P. P.; Ferreira, M. A. Discrimination between Arabica and Robusta coffee species on the basis of their amino acid enantiomers. J. Agric. Food Chem. **2003**, *51*, 6495–6501.

(7) Martin, M. J.; Pablos, F.; Gonzalez, A. G.; Valdenebro, M. S.; Leon-Camacho, M. Fatty acid profiles as discriminant parameters for coffee varieties differentiation. Talanta **2001**, *54*, 291–297.

(8) Ky, C. L.; Louarn, J.; Dussert, S.; Guyot, B.; Hamon, S.; Noirot, M. Caffeine, trigonelline, chlorogenic acids and sucrose diversity in wild *Coffea arabica L.* and *C. canephora P.* accessions. Food Chem. **2001**, *75*, 223–230.

(9) Andrade, P. B.; Leitao, R.; Seabra, R. M.; Oliveira, M. B.; Ferreira, M. A. 3,4-Dimethoxycinnamic acid levels as a tool for differentiation of *Coffea canephora* var. robusta and *Coffea arabica*. Food Chem. **1998**, *61*, 511–514.

(10) Huck, C. W.; Guggenbichler, W.; Bonn, G. K. Analysis of caffeine, theobromine and theophylline in coffee by near infrared spectroscopy (NIRS) compared to high-performance liquid chromatography (HPLC) coupled to mass spectrometry. Anal. Chim. Acta **2005**, *538*, 195–203.

(11) Gonzalez, A. G.; Pablos, F.; Martin, M. J.; Leon-Camacho, M.; Valdenebro, M. S. HPLC analysis of tocopherols and triglycerides in coffee and their use as authentication parameters. Food Chem. **2001**, *73*, 93–101.

(12) Carrera, F.; Leon-Camacho, M.; Pablos, F.; Gonzalez, A. G. Authentication of green coffee varieties according to their sterolic profile. Anal. Chim. Acta **1998**, *370*, 131–139.

(13) Gil-Agusti, M. T.; Campostrini, N.; Zolla, L.; Ciambella, C.; Invernizzi, C.; Righetti, P. G. Two-dimensional mapping as a tool for classification of green coffee bean species. Proteomics **2005**, *5*, 710–718.

(14) Bertrand, B.; Vaast, P.; Alpizar, E.; Etienne, H.; Davrieux, F.; Charmetant, P. Comparison of bean biochemical composition and beverage quality of Arabica hybrids involving Sudanese−Ethiopian origins with traditional varieties at various elevations in Central America. Tree Physiol. **2006**, *26*, 1239–1248.

(15) Rubayiza, A. B.; Meurens, M. Chemical discrimination of arabica and robusta coffees by Fourier transform raman spectroscopy. J. Agric. Food Chem. **2005**, *53*, 4654–4659.

(16) Clifford, M. N.; Jarvis, T. The chlorogenic acids content of green robusta coffee beans as a possible index of geographic origin. Food Chem. **1988**, *29*, 291–298.

(17) Clifford, M. N.; Knight, S. The cinnamoyl-amino acid conjugates of green robusta coffee beans. Food Chem. **2004**, *87*, 457–463.

(18) Serra, F.; Guillou, C. G.; Reniero, F.; Ballarin, L.; Cantagallo, M. I.; Wieser, M.; Iyer, S. S.; Heberger, K.; Vanhaecke, F. Determination of the geographical origin of green coffee by principal component analysis of carbon, nitrogen and boron stable isotope ratios. *Rapid Commun. Mass Spectrom.* **2005**, *19*, 2111–2115.

(19) Weckerle, B.; Richling, E.; Heinrich, S.; Schreier, P. Origin assessment of green coffee (*Coffea arabica*) by multi-element stable isotope analysis of caffeine. Anal. Bioanal. Chem. **2002**, *374*, 886–890.

(20) Clifford, M. N.; Johnston, K. L.; Knight, S.; Kuhnert, N. Hierarchical scheme for LC-MS^n identification of chlorogenic acids. J. Agric. Food Chem. **2003**, *51*, 2900–2911.

(21) Clifford, M. N.; Knight, S.; Kuhnert, N. Discriminating between the six isomers of dicaffeoylquinic acid by LC-MS^n. J. Agric. Food Chem. **2005**, *53*, 3821–3832.

(22) Clifford, M. N.; Knight, S.; Surucu, B.; Kuhnert, N. Characterization by LC-MS^n of four new classes of chlorogenic acids in green coffee beans: dimethoxycinnamoylquinic acids, diferuloylquinic acids, caffeoyl-dimethoxycinnamoylquinic acids, and feruloyl-dimethoxycinnamoylquinic acids. J. Agric. Food Chem. **2006**, *54*, 1957–1969.

(23) Clifford, M. N.; Marks, S.; Knight, S.; Kuhnert, N. Characterization by LC-MS^n of four new classes of *p*-coumaric acid-containing diacyl chlorogenic acids in green coffee beans. J. Agric. Food Chem. **2006**, *54*, 4095–4101.

(24) Alonso-Salces, R. M.; Guillou, C.; Berrueta, L. A. Liquid chromatography coupled with ultraviolet absorbance detection, electrospray ionisation, collision-induced dissociation and tandem mass spectrometry for the on-line characterisation of polyphenols and methylxanthines in green coffee beans. Rapid Commun. Mass Spectrom. **2009**, *23*, 363–383.

(25) Clifford, M. N. Chlorogenic acids and other cinnamates—nature, occurrence, dietary burden, absorption and metabolism. J. Sci. Food Agric. **2000**, *80*, 1033–1043.

(26) Farah, A.; Monteiro, M. C.; Calado, V.; Franca, A. S.; Trugo, L. C. Correlation between cup quality and chemical attributes of Brazilian coffee. Food Chem. **2006**, *98*, 373–380.

(27) Ky, C. L.; Louarn, J.; Guyot, B.; Charrier, A.; Hamon, S.; Noirot, M. Relations between and inheritance of chlorogenic acid contents in an interspecific cross between *Coffea pseudozanguebariae* and *Coffea liberica* var. 'dewevrei'. Theor. Appl. Genet. **1999**, *98*, 628–637.

(28) Clifford, M. N. Chlorogenic acids. In *Coffee. Vol. 1: Chemistry*; Clarke, R. J., Macrae, R., Eds.; Elsevier Applied Science: London, U.K., 1985; pp 153−202.

(29) De Menezes, H. C. The relationship between the state of maturity of raw coffee beans and the isomers of caffeoylquinic acid. Food Chem. **1994**, *50*, 293–296.

(30) Clifford, M. N.; Kazi, T. The influence of coffee bean maturity on the content of chlorogenic acids, caffeine and trigonelline. Food Chem. **1987**, *26*, 59–69.

(31) Vaast, P.; Bertrand, B.; Perriot, J. J.; Guyot, B.; Genard, M. Fruit thinning and shade improve bean characteristics and beverage quality of coffee (*Coffea arabica* L.) under optimal conditions. J. Sci. Food Agric. **2006**, *86*, 197–204.

(32) Berrueta, L. A.; Alonso-Salces, R. M.; Héberger, K. Supervised pattern recognition in food analysis. J. Chromatogr., A **2007**, *1158*, 196–214.

(33) IUPAC, IUPAC Commission on the Nomenclature of Organic Chemistry (CNOC) and IUPAC-IUB Commission on Biochemical Nomenclature (CBN). Nomenclature of cyclitols. Recommendations, 1973. Biochem. J. 1976, *153*, 23−31.

(34) Correia, A. M. N. G.; Leitao, M. C. A.; Clifford, M. N. Caffeoyl-tyrosine and Angola II as characteristic markers for Angolan robusta coffees. Food Chem. **1995**, *53*, 309–313.

(35) Macheix, J. J.; Fleuriet, A.; Billot, J. *Fruit Phenolics*; CRC Press: Boca Raton, FL, 1990.

(36) Anthony, F.; Clifford, M. N.; Noirot, M. Biochemical diversity in the genus *Coffea L.*: chlorogenic acids, caffeine and mozambioside contents. Genet. Resour. Crop Evol. **1993**, *40*, 61–70.

(37) Perrone, D.; Farah, A.; Donangelo, C. M.; de Paulis, T.; Martin, P. R. Comprehensive analysis of major and minor chlorogenic acids and lactones in economically relevant Brazilian coffee cultivars. Food Chem. **2008**, *106*, 859–867.

(38) Balyaya, K. J.; Clifford, M. N. Individual chlorogenic acids and caffeine contents in commercial grades of wet and dry processed Indian green robusta coffee beans. J. Food Sci. Technol. Mysore **1995**, *32*, 104–108.

(39) Sharaf, M.; llman, D.; Kowalski, B. R. *Chemometrics*; Wiley: New York, 1986.

(40) Esbensen, K. H.; Guyot, D.; Westad, F.; Houmøller, L. P. PLS-Regression (PLS-R). In *Multivariate Data Analysis—in Practice: An Introduction to Multivariate Data Analysis and Experimental Design*, 5th ed.; Esbensen, K. H., Ed.; Camo Process: Oslo, Norway, 2006; pp 137−154.

Received for Review November 28, 2008. Revised manuscript received February 26, 2009. Accepted February 28, 2009.