

# Variation in bean morphology and biochemical composition measured in different genetic groups of arabica coffee (*Coffea arabica* L.)

Hue T. M. Tran<sup>1,2</sup> · Carlos Alberto Cordero Vargas<sup>3</sup> · L. Slade Lee<sup>1,4</sup> · Agnelo Furtado<sup>1</sup> · Heather Smyth<sup>1</sup> · Robert Henry<sup>1</sup>

Received: 11 August 2016 / Revised: 22 March 2017 / Accepted: 5 April 2017 / Published online: 13 April 2017  
© Springer-Verlag Berlin Heidelberg 2017

**Abstract** The narrow genetic base of commercial arabica resulting from intensive selection for quality during domestication and self-pollination has been well documented, raising the need for new diverse germplasm sources. Beans of 232 diverse arabica coffee accessions originating from 27 countries were harvested from the germplasm collection at CATIE, Costa Rica. Substantial variation was observed for bean morphology including 100 bean weight, bean length, width, thickness and bulk density. Non-volatiles including caffeine and trigonelline were analysed and showed larger variation in range than has previously been reported. Results of targeted analysis of 18 volatiles from 35 accessions also showed significant variation, with coefficients of variation from 140% for

4-vinylguaiaicol to 62% for geraniol. There were strong correlations between some volatile compounds, suggesting that representative volatiles used in selection would save analytical costs. However, no strong correlation was found between bean morphology and the levels of non-volatile or volatile compounds, implying that it is difficult to select for low or high composition of these compounds based on bean physical characteristics. Utilizing the large variation observed for bean morphology and biochemical traits, it should be possible to select for desirable combinations of traits in arabica coffee breeding.

**Keywords** Arabica coffee · Coffee bean · Volatile compounds · Non-volatile compounds · Caffeine · Trigonelline

**Key message:** There is substantial variation for bean morphology, non-volatile and volatile compounds in arabica coffee germplasm. These traits have no significant correlation indicating the potential for independent selection.

Communicated by A. M. Dandekar

**Electronic supplementary material** The online version of this article (doi:10.1007/s11295-017-1138-8) contains supplementary material, which is available to authorized users.

✉ Hue T. M. Tran  
t.tran22@uq.edu.au

✉ Robert Henry  
robert.henry@uq.edu.au

<sup>1</sup> Queensland Alliance for Agriculture and Food Innovation (QAAFI), The University of Queensland, St. Lucia, QLD 4072, Australia

<sup>2</sup> Western Highlands Agriculture & Forestry Science Institute (WASI), Daklak, Vietnam

<sup>3</sup> Centro Agronómico Tropical de Investigación y Enseñanza (CATIE), Turrialba 30501, Costa Rica

<sup>4</sup> Southern Cross University, Lismore, NSW 2480, Australia

## Introduction

Coffee belongs to the Rubiaceae family, the *Coffeae* tribe and the *Coffea* genus with more than 124 species (Davis et al. 2011). Commercial coffee production is dominated by only two species namely *Coffea arabica* L. (known as arabica coffee) and *Coffea canephora* P. (generally referred to as robusta coffee). All coffee species are diploid ( $2n = 2x = 22$ ) and generally self-incompatible, except for *C. arabica* which is the only tetraploid ( $2n = 4x = 44$ ), self-fertile, derived from a spontaneous hybridization between *C. canephora* (as paternal progenitor) and *Coffea eugenioides* (as maternal progenitor) and known to have a very narrow genetic base (Charrier and Eskes 2004). There are two genetic origins of arabica: Typica and Bourbon from which all of the commercial varieties in the world today are derived (Anthony et al. 2001). The low genetic diversity among cultivars is probably due to the narrow geographical origin, the selection bottleneck and self-pollination (Teressa et al. 2010; Tran 2005). A number of studies have focussed on assessment of arabica

diversity using diverse groups of materials such as cultivar, mutants, spontaneous/subspontaneous accessions (terms defined in Charrier and Eskes 2004) and hybrids as well as different marker systems (Aerts et al. 2012; Aga et al. 2003; Dessalegn et al. 2008b). These studies indicated a limited genetic variability available for arabica breeding programs, even with the wild accessions (Aga et al. 2003; Anthony et al. 2001; WCR 2014). Further assessment in a wider germplasm collection is required to identify genetically diverse material for use as genetic sources for improving arabica coffee quality.

Coffee quality is assessed based on both physical and organoleptic quality. The physical quality characteristics include length, width, thickness or weight, shape or colour of coffee beans, while the organoleptic quality relates to the fragrance or aroma, flavour, sweetness, acidity or overall taste (Leroy et al. 2006). Both the physical and chemical attributes of coffee beans are important criteria which determine the value of arabica coffee beans in world markets (Belete 2014).

Caffeine and trigonelline are key non-volatile compounds in coffee. Caffeine is one of the main alkaloids contributing to the strength, body and bitterness of brewed coffee (Trugo 1984). Trigonelline is a methyl-betaine biologically derived from enzymatic methylation of nicotinic acid (Farah 2009). The levels of trigonelline in green and roasted coffee beans are strongly correlated with high quality (Farah et al. 2006b). Trigonelline contributes indirectly to the formation of different desirable aromas during the roasting process (Ky et al. 2001b), and thus, increasing trigonelline is also a breeding objective.

Coffee flavour and its formation are extremely complex; however, a number of studies on coffee flavour and its constituents have been reported. The contribution of volatile and non-volatile compounds to coffee quality and their relationship to coffee sensory properties has been thoroughly reviewed by Flament (2002) and more recently by Sunarharum et al. (2014). The reactions involved in the formation of coffee aroma including the Maillard reaction (non-enzymatic browning), the Strecker degradation, the breakdown of sulphur amino acids, hydroxy-amino acids and proline or the degradation of trigonelline, quinic acid moieties and pigments and its mechanisms have been reviewed by Buffo and Cardelli-Freire (2004). Among 1000 volatile compounds known in roasted bean, 20 were identified as key aroma compounds (Oestreich-Janzen 2010) and belong to several chemical classes including pyrazines, ketones and aldehydes (Flament 2002).

There are many analytical methods available for the determination of caffeine and trigonelline in coffee. However, high-performance liquid chromatography (HPLC) has been a common technique due to its accuracy, precision and high throughput (De Maria et al. 1995). The analysis of caffeine and trigonelline in coffee by HPLC was reported by a number of authors (Casal et al. 1998; Casal et al. 2000; De Maria et al. 1995; Trugo and Macrae 1989). The extraction methods,

analytical and detection systems have been thoroughly reviewed (Belay 2011; Jeszka-Skowron et al. 2015; Nuhu 2014). The quantification of volatile compounds is more difficult. Stable isotope dilution analysis (SIDA) combined with head space-solid-phase microextraction (HS-SPME) method/gas chromatography-mass spectrometry (GC-MS) has recently been applied to coffee quality analysis with advantages of high repeatability, sensitivity, automation, speed of analysis and avoiding the drawbacks related to the matrix effect (Bicchi et al. 2011; Pickard et al. 2013).

In this study, we report the morphometric and chemical variation of a *C. arabica* population collected from one of the world's largest coffee collections based on measurement of bean physical properties and composition using state-of-the-art analytical technologies. The study targets outcomes to help guide breeding programs in selection for favourable bean size, caffeine, trigonelline and volatile compounds.

## Materials and methods

### Materials

Beans of 235 coffee accessions originated from 27 countries from the germplasm collection of 1730 accessions conserved at Centro Agronómico Tropical de Investigación y Enseñanza (CATIE, Costa Rica) were selected for this study. The collection is located in an area of approximately of 8.5 ha at latitude 9° 53' 57" N, at longitude 83° 39' 43" W and at an elevation of 616.1 m above sea level. It has an average annual rainfall of 2779 mm and an average annual temperature of 22.2 °C. Its soil has a pH of 4.76 and contains 3.62% organic carbon, 0.32% total nitrogen and 11.3 mg/l phosphorus (provided by CATIE, obtained in 2012). Most of the accessions were grown from seed, each of which constitutes a genotype. Each accession has four to eight trees. The spacing between rows and between trees within rows is generally 2.5 m × 2 m. *Erythrina popeppigiana* trees were grown as shade trees at the spacing of 6 m × 6 m. The collection's maintenance was described in detail by Anthony et al. (2007a). Harvesting is conducted four times in a year from July to November.

The selection of this germplasm set was based on its representation of the core collection, accessions known for biochemical compounds or quality, country diversity, priority given to wild-type Ethiopian accessions and hybrids and availability of fruits at the collection time. Accessions were classified into three groups based on their genetic origins: (1) accessions from the diversity centre of *C. arabica* (wild type), (2) accessions from Typica and Bourbon-derived varieties and mutants and (3) accessions from introgressed lines selected from interspecific hybrid progeny (*C. arabica* × *Coffea* spp.) (Anthony et al. 2007b). Out of 235 accessions, 232 were *C. arabica*, 2 were *C. canephora* and 1 was *C. brevipes*

(Supplement Table 1). These two species were used as reference due to their known high caffeine and low trigonelline (Campa et al. 2004; Campa et al. 2005a; Ky et al. 2001a).

## Methods

### *Sample collection, processing and storage*

Fruits were collected from one single tree per accession in November 2014. Ripe cherry was harvested by hand and then taken to a processing house for pulping. After removing the pulp, fresh coffee bean was stored in nylon (PE) bags with water (due to the small amount of coffee bean) for the removal of mucilage. After 12 h, the fresh coffee bean was taken out of the plastic bag and washed with clean water. Then, the fresh cleaned bean was dried in the shade and open air for 1–2 weeks. The parchment was then de-husked. The bean was measured for the moisture content using a MD1229 SEEDBURO 1200S at 12 °C, then placed in a hermetic box and shipped to the University of Queensland (UQ), Australia. Bean was stored in the refrigerator at 4 °C prior to measurement, roasting and chemical analysis.

### *Bean physical measurement*

Bean physical characteristics were measured by applying the method developed by IPGRI (1996) with modification for the weight of 100 beans (W100) as follows: weight of 100 beans at 12% moisture content =  $[(1 - x\%) / (1 - 12\%)] \times$  weight of 100 beans calculated at  $x\%$  moisture, measured in triplicate. Bean length, width and thickness were measured for five replicates using a calliper. The bulk density of green bean was measured in duplicate by weighing the dried beans in a 100-ml cylinder and then converting to kilogram per cubic metre.

### *Non-volatile analysis*

**Sample extraction** Samples were extracted for non-volatile analysis using a method based on that of Casal et al. (1998) with a minor modification. Two grammes of green coffee powder of each sample was placed in a 100-ml Erlenmeyer flask with a magnetic stirrer and boiled in 20 ml of MilliQ water for 2.5 min and then allowed to cool for about 2.5 min before transferring the extract into a 100-ml volumetric flask. The boiling cycle was repeated two more times and extracts mixed together in the volumetric flask and diluted to a final volume of 100 ml with MilliQ water. The extract was transferred into a 50-ml tube and then cooled down by placing in a -20 °C freezer for 12 min. The cooled extract was first centrifuged at 1500 rpm (252×g) for 5 min at 4 °C, and the supernatant filtered using a 0.13- $\mu$ m syringe filter. The filtered extracts were stored in HPLC vials at -20 °C for further

analysis. Analyses were conducted in duplicate with separated bean samples.

**Instrumental conditions** Samples were analysed on a HPLC/diode array detector system. The HPLC conditions followed those of Casal et al. (2000) using a Spherisorb S5 ODS2 (0.46 × 25.0 cm) column, with a  $\mu$ Bondapak C18 (10  $\mu$ m) guard column (Waters) and a Shimadzu chromatograph (model LC-10 VP). The temperature of the column and the autosampler was 23 and 4 °C, respectively. The injection volume was 20  $\mu$ l with a gradient as follows: (a) phosphate buffer (KH<sub>2</sub>PO<sub>4</sub>) 0.1 M (pH 4.0) and (b) methanol at 0 min (7% B), 4 min (9% B), 6 min (25% B), 13 min (29% B), 21 min (50% B) and 30 min (7% B) at a flow rate of 1 ml/min. Detection in both cases was achieved using a diode array detector at 265 nm for trigonelline and 273 nm for caffeine. The compounds under study were identified by chromatographic comparisons with authentic standards (caffeine and trigonelline hydrochloride standards were from Sigma-Aldrich, Sydney, Australia). Analytical grade methanol and KH<sub>2</sub>PO<sub>4</sub> were also obtained from Sigma-Aldrich (Sydney, Australia).

Quantification was based on an external standard curves. Four calibration points were used to create a linear calibration curve in order to quantify the compounds (caffeine 0.05–500  $\mu$ g/ml, trigonelline 0.15–450  $\mu$ g/ml). The correlation coefficient between the external standard concentrations and absorbance values for each standard curve invariably exceeded 0.999.

### *Coffee bean roasting*

Coffee beans were roasted at Ashtan Place (Banyo, QLD 4014) by an experienced coffee roaster, Peter Wolff. Since the cherry samples were harvested from single trees at one harvest time, the quantity available was very limited with a sample size ranging from 24 to 188 g. Among the 235 samples collected, 221 samples for which at least 50 g of bean was available were roasted. This sample size was chosen based on preliminary experiments comparing different roasting sample sizes at 25, 50, 75 and 100 g using the Red Catuai variety from Brazil and assessing for aroma and volatile compounds using GC-MS. Since samples had varying bean size, each sample was roasted until a consistent colour was achieved. Each sample was roasted at start temperature of 180 °C; then, the temperature was increased to 188 °C and then adjusted to 185 °C for 6–7 min total time and then cooled for 4 min. The roasting machine used was designed for research scale use and was a Coffee PRO, sample Pro 100 Gas, capacity 2 × 100 g, date of production of December 2010 (Nexu International Ltd., [coffeetrays.com](http://coffeetrays.com)). The colour of the roasted bean was light to medium as measured using a Roast Analyser—RoAmi, Roaster's Friend (TRUE systems, [true-systems.com](http://true-systems.com)). This colour range was based on SCAA/Agtron (Agtron 2004).

Almost all samples (97.73%) were roasted from very light to medium light indicating that a comparable roasting of the samples had been achieved (supplementary Table 2).

#### Targeted analysis for volatiles

Two hundred twenty-one samples were subjected to full-scan non-targeted analysis (HS-SPME/GC-MS). Their chromatographic fingerprints (total ion chromatographs (TICs)) were plotted in a principal component analysis (PCA) from which 35 samples showed contrasting characters for principal components 1 and 2 with some outlying wild-type genotypes that were selected for targeted analysis (supplementary Fig. 1).

The targeted method involved SIDA together with HS SPME/GC-MS for 18 important volatile flavour compounds. The details of the extraction method, instrumental conditions and calibration parameters of the method applied are described by Sunarharum (2016).

**Sample and standard preparation and extraction** The 35 roasted bean samples were stored at  $-20^{\circ}\text{C}$  until prepared for analysis. Samples were ground using a Retsch Mixer Mill 4000 with a frequency of 30/s for 30 s and weighed (2 g) in triplicate into a screw cap glass sample vial together with 2 ml of saturated brine (NaCl), 2 ml milliQ water and a magnetic stirrer flea. Similarly, spiked standard addition calibration solutions were prepared and weighed (2 g) into vials together with brine, water and a stir flea. A spike of internal standard solution was added which contained isotope analogues of the target volatiles. Sample extraction was conducted using a Gerstel MPS2 Autosampler (Gerstel, Germany). SPME fibre type, extraction time and temperature were as described by Sunarharum (2016).

#### Development of calibrations and instrumental conditions

Instrumental analysis was performed using an Agilent 6890N gas chromatograph (GC) and an Agilent 5975 series mass selective detector (MSD) unit (Agilent Technologies Inc., CA, USA) with a MPS-2 multipurpose sampler (Gerstel, Germany) installed according to the instrumental parameters and conditions described by Sunarharum (2016). All analyses were performed in triplicate. A four-point calibration function was developed for each target compound.

Quantitation was achieved using selected ion monitoring, and qualifying ions were used to identify target compounds together with matching retention times with authentic reference standards. Co-elution with other compounds was common as coffee has a great number of compounds; however, there was no co-elution of target compound ions used for quantification. The data were collected using the Enhanced ChemStation software MSD ChemStation G1701EA revision E.02.02, and compound concentration data were exported into excel prior to analysis.

#### Data analysis

One-way analysis of variance (ANOVA) was performed using the software GenStat version 11 (Payne et al. 2008) using category of accessions (cultivars, hybrids and wild) as a factor. Mean values between two groups were compared using linear contrasts. Spearman rank correlations were used to examine the relationship between variables after testing for normality using Shapiro-Wilk test.

For each variable, the value of measurement for each individual was standardized by subtracting it from the population mean and dividing by the standard deviation in order to reduce the influence of the scale differences. The standardized data were then used in PCA performed with the software XLSTAT (Addinsoft 2007) (for targeted analysis in which 18 variables (18 volatile compounds) and 35 samples were included). For non-targeted analysis (supplementary data), due to the large data set (100 selected chromatographic variables and 221 samples), PCA multivariate exploration was conducted using the Unscrambler X, version 10.3 (Martens et al. 1987).

## Results

### Variation in bean physical characteristics

Measurement of 232 arabica accessions for green bean physical characteristics showed that 100 bean weight (W100), length, width, ratio between length and width, thickness and bulk density were on average  $15.53 \pm 0.12$  g,  $9.62 \pm 0.08$  mm,  $6.83 \pm 0.05$  mm,  $1.41 \pm 0.08$ ,  $3.95 \pm 0.07$  mm and  $600 \pm 0.05$  kg/m<sup>3</sup>, respectively. W100 and bean length had the largest variation, with the coefficient of variation (CV) of 12.20 and 7.73%, respectively (Table 1). There was a significant difference ( $P < 0.001$ ) between accessions for all variables. Although the three arabica groups had similar mean weights of 100 beans (15.35, 15.80 and 15.57 g for cultivars, hybrids and wild type, respectively), the hybrid group comprising both intraspecific and interspecific hybrids had the largest variation (CV = 15.54%), followed by the cultivar group (CV = 12.50%), and the least variable was the wild-type group (CV = 10.69%) even though this group accounts for almost half of the population. This group was also the least variable for length (6.99 vs 8.33 and 8.34%) and width (5.01 vs 5.27 and 6.13%). For thickness and bulk density, hybrids had the least variability. However, the difference between groups was not significant for W100, length and bulk density. For the ratio between length and width, width and thickness, the wild-type group was significantly different to the cultivar and hybrid groups ( $P < 0.05$ ), while there was no significant difference between the cultivar and hybrid groups. The wild group had slimmer beans and a greater thickness than the cultivars and hybrids (Table 1).

**Table 1** Variation in green bean physical quality measured in 232 arabica coffee accessions

Variable	Min	Max	Mean	CV (%)	LSD( $P < 0.05$ )	Literature
W100 (g)	10.16	23.13	15.53	12.20	0.54	9.77–21.82 <sup>a</sup>
+ Cultivar ( $n = 83$ )	10.16	23.13	15.35a	12.50		
+ Hybrids ( $n = 36$ )	12.17	22.71	15.80a	15.54		
+ Wild type ( $n = 113$ )	11.87	19.88	15.57a	10.69		
Length (L) (mm)	6.64	13.20	9.62	7.73	1.12	8.19–11.04 <sup>b</sup>
+ Cultivar ( $n = 83$ )	6.64	13.20	9.50a	8.33		
+ Hybrids ( $n = 36$ )	8.32	12.52	9.75a	8.34		
+ Wild type ( $n = 113$ )	8.01	12.05	9.66a	6.99		
Width (W) (mm)	5.97	8.08	6.83	5.43	0.59	6.11–8.27 <sup>b</sup>
+ Cultivar ( $n = 83$ )	5.97	8.07	6.91a	5.27		
+ Hybrids ( $n = 36$ )	6.22	7.92	6.94a	6.13		
+ Wild type ( $n = 113$ )	6.12	8.08	6.74b	5.01		
Ratio L/W	1.11	1.75	1.41	7.68	0.15	1.33–1.35 <sup>b</sup>
+ Cultivar ( $n = 83$ )	1.11	1.67	1.38b	7.13		
+ Hybrids ( $n = 36$ )	1.18	1.65	1.41b	7.58		
+ Wild type ( $n = 113$ )	1.22	1.75	1.44a	7.45		
Thickness (mm)	3.27	5.18	3.95	7.04	0.49	4.60–5.13 <sup>b</sup>
+ Cultivar ( $n = 83$ )	3.27	4.84	3.88b	7.54		
+ Hybrids ( $n = 36$ )	3.41	4.49	3.87b	5.80		
+ Wild type ( $n = 113$ )	3.42	5.18	4.01a	6.66		
Bulk density (kg/m <sup>3</sup> )	454	679	600	4.57	20.86	635–707 <sup>c</sup>
+ Cultivar ( $n = 83$ )	458	675	603a	5.24		
+ Hybrids ( $n = 36$ )	566	664	605a	3.76		
+ Wild type ( $n = 113$ )	505	679	597a	4.54		

Values with the same alphabet along the same column are not significantly different ( $P < 0.05$ )

CV (%) coefficient of variation, LSD( $P < 0.05$ ) least significant difference at  $P$  value less than 0.05

<sup>a</sup> Belete (2014), Bertrand et al. (2005) and Montagnon and Bouharmont 1996

<sup>b</sup> Olukunle and Akinnuli (2012) and Ghosh and Gacanja (1970)

<sup>c</sup> Franca et al. (2005) and Rodrigues et al. (2009)

### Variation in non-volatile compounds

The green bean caffeine content of the 232 arabica accessions was on average 1.25% on a dry matter basis (d.m.b), ranging from 0.82 to 1.76%, while the trigonelline content was 1.11% on average, ranging from 0.80 to 1.38%. The coefficient of variation was 9.64% for caffeine and 9.55% for trigonelline (Table 2). Significant genotypic variation ( $P < 0.001$ ) in both caffeine and trigonelline contents was observed among accessions (Table 2); they appeared to follow a normal distribution that was slightly skewed towards the lower value (Fig. 1). The cultivar group had a significantly ( $P < 0.05$ ) lower caffeine content than the hybrid and wild groups (Table 2). With regards to within-group variation, the hybrid group was the most diverse in caffeine (1.01–1.76%, CV = 11.29%) despite having the smallest number of accessions (38), followed by the cultivar group (0.82–1.52%, CV = 9.18%) and the wild-type group from Ethiopia (1.18–1.51%, CV = 9.15%)

(Table 2). For trigonelline, however, there was no significant difference ( $P < 0.05$ ) between the three groups.

### Variation in volatile compounds

A preliminary principal component analysis of the chromatographic fingerprints derived from the non-targeted scanning of 221 samples allowed identification of 35 accessions with contrasting distribution on the two axis and some outlying wild-type genotypes (Supplementary Fig. 1). Targeted analysis of these accessions showed significant variation ( $P < 0.001$ ), with CV% varying from 14% (for 4-vinylguaiacol) to 62% (for geraniol). The volatile concentration ranged from 9 ppb for beta damascenone to over 53,300 ppb for 2,5-dimethylpyrazine, and all volatiles quantified were far above their reported sensory threshold concentration which verifies their importance as target aroma volatiles in arabica (Table 3). Five of the 18 compounds (2,3-pentanedione, 3-

**Table 2** Variation in caffeine and trigonelline content among different arabica coffee groups

Species	Group	Types	Size	Min	Max	Mean	CV%	LSD( $P < 0.05$ )
Caffeine (% d.m.b.)								
<i>C. arabica</i>	1	Variety/selection	83	0.82	1.52	1.23b	9.18	0.13
	2	Hybrid	36	1.01	1.76	1.28a	11.29	0.07
	3	Wild type	113	1.18	1.51	1.26a	9.15	0.11
	Total/average			232	0.82	1.76	1.25	9.64
<i>C. canephora</i>			2	1.19	2.48	1.84		
<i>C. brevipes</i>			1			2.37		
Trigonelline (% d.m.b.)								
<i>C. arabica</i>	1	Variety/selection	83	0.89	1.38	1.12a	9.66	0.11
	2	Hybrid	36	0.89	1.31	1.13a	9.69	0.10
	3	Wild type	113	0.94	1.31	1.10a	9.46	0.10
	Total/average			232	0.80	1.38	1.11	9.55
<i>C. canephora</i>			2	0.91	1.24	1.08		
<i>C. brevipes</i>			1			0.76		

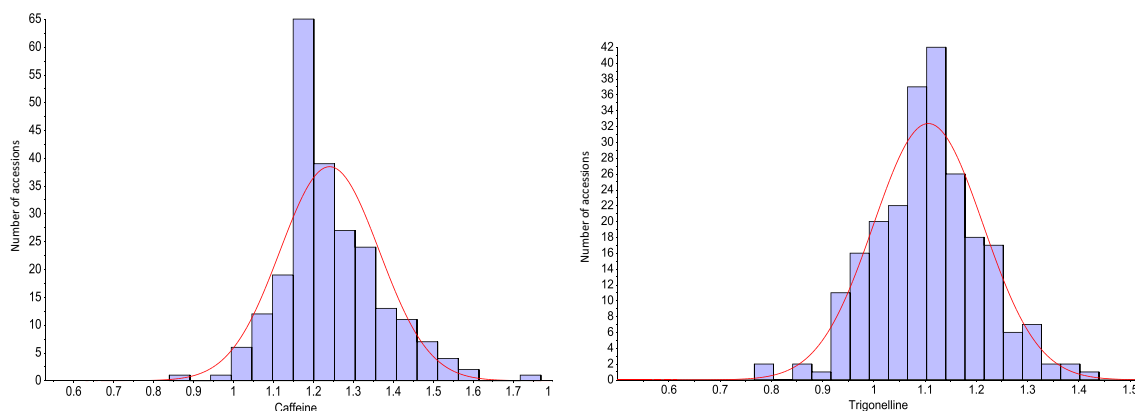
Values with the same alphabet along the same column are not significantly different ( $P < 0.05$ )

CV (%) coefficient of variation, LSD( $P < 0.05$ ) least significant difference at  $P$  value less than 0.05

methylbutanal, 2-methylbutanal, 4-vinylguaiacol and 2,3-diethyl-5-methylpyrazine) had concentrations in the range previously reported. Eight compounds (2,3-butanedione, 3-methylbutanal, guaiacol, 4-ethylguaiacol and beta damascenone, 3-isobutyl-2-methoxypyrazine, linalool, limonene) had concentrations lower than those reported in the literature. Four compounds belonging to the pyrazines class (2,5-dimethylpyrazine, 2-ethyl-3,6-dimethylpyrazine, 2,3-dimethylpyrazine and 2-ethyl-3,5-dimethylpyrazine) and aldehydes (2-methylpropanal) had concentrations far exceeding those reported in literature (Belitz et al. 2009; Cheong et al. 2013; Czerny et al. 1999; Piccino et al. 2014). For geraniol, no published data were available for comparison.

Principal component analysis of 18 volatiles measured in 35 accessions showed that the first two components PC1 and PC2 explained 61% of the total variation (Fig. 2). The population was distributed along the PC1

with contrasting volatile profile, either low or high in almost all volatiles, but mostly affected by pyrazines (2,5-dimethylpyrazine, 2-ethyl-3,5-dimethylpyrazine and 2-ethyl-3,6-dimethylpyrazine) and aldehydes (2-methylpropanal, 3-methylbutanal and 2-methylbutanal) (Fig. 2). PC2 showed coffee accessions that were either higher in phenolic compounds (guaiacol, 4-ethylguaiacol) and lower in linalool (terpenes) and 3-isobutyl-2-methoxypyrazine (pyrazines) or vice versa. Almost all volatile compounds of the same chemical class were clustered into groups such as aldehydes, phenolic compounds, pyrazines and ketones. Some accessions showed high concentrations in most compounds (e.g. 4857, 4909 and 4918), while others (e.g. 4288, 3483 and 3747) were low in most compounds. Accessions 1993 and 2268 were highest in linalool (terpenes) and 3-isobutyl-2-methoxypyrazine (pyrazines) and lowest in phenolic



**Fig. 1** The distribution of caffeine (left) and trigonelline (right) content in the arabica coffee population ( $n = 232$ )

**Table 3** Variation in 18 key volatile compounds measured in a representation of 35 arabica coffee accessions selected from non-targeted analysis

Compounds	Min (ppb)	Max (ppb)	Mean (ppb)	% CV	LSD ( $P < 0.05$ )	Literature (ppb)	Sensory threshold (ppb)	Aroma description
<b>Ketones</b>								
2,3-Butanedione	1426	9158	4255	47.80	3158	48,400–50,800 <sup>a</sup>	0.3 <sup>a</sup>	Buttery oily <sup>a</sup> , fruity, caramel-like <sup>b,d</sup>
2,3-Pentanedione	2256	27,932	12,258	47.76	4330	3540–39,600 <sup>a</sup>	20 <sup>a</sup>	Buttery oily <sup>a</sup> , caramel-like <sup>b,d</sup>
<b>Aldehydes</b>								
2-Methylpropanal	1050	4773	2344	36.23	859	320–430 <sup>c</sup>	3 <sup>c</sup>	Buttery oily <sup>a</sup>
3-Methylbutanal	896	3723	2014	32.79	541	210–18,600 <sup>a,c</sup>	0.35 <sup>a</sup>	Malty <sup>a,c</sup> , chocolate <sup>c</sup> , caramel-like <sup>d</sup>
2-Methylbutanal	257	1052	523	36.01	169	200–20,700 <sup>a,c</sup>	1.3 <sup>a</sup>	Chocolate <sup>c</sup> , caramel-like <sup>d</sup>
<b>Phenolic compounds</b>								
Guaiacol	381	3384	1479	42.80	504	2000–17,970 <sup>a</sup>	2.5 <sup>a</sup>	Phenolic, burnt <sup>a,b</sup> , smoky, phenolic <sup>d</sup>
4-Ethylguaiacol	64	371	186	44.40	63	800–24,800 <sup>a</sup>	25 <sup>a</sup>	Spicy <sup>a</sup> , sweet <sup>b</sup> , smoky, phenolic <sup>d</sup>
4-Vinylguaiacol	22,723	38,773	30,253	13.50	4818	8000–64,800 <sup>a</sup>	0.75 <sup>a</sup>	Spicy <sup>a</sup> , clove-like <sup>b</sup> , smoky, phenolic <sup>d</sup>
<b>Norisoprenoids</b>								
Beta damascenone	7	14	9	18.96	2	195–255 <sup>a</sup>	0.00075 <sup>a</sup>	Honey-line, fruity <sup>a,d</sup>
<b>Pyrazines</b>								
2,5-Dimethylpyrazine	19,201	123,612	53,312	38.15	19,484	4550–14,070 <sup>a,c</sup>	80 <sup>a</sup>	–
2-Ethyl-3,6-dimethylpyrazine	2685	12,320	6390	35.76	20	2570–5980 <sup>a</sup>	8.6 <sup>a</sup>	–
2,3-Diethyl-5-methylpyrazine	41	134	75	30.69	25	73–95 <sup>a</sup>	0.09 <sup>a</sup>	Nutty-roast <sup>a</sup> , earthy <sup>d</sup>
2,3-Dimethylpyrazine	3131	33,662	11,963	50.05	10,192	2580–8890 <sup>a,c</sup>	800 <sup>a</sup>	–
2-Ethyl-3,5-dimethylpyrazine	479	2136	996	33.60	326	55–840 <sup>a,c</sup>	0.04 <sup>a</sup>	Nutty-roast <sup>a</sup> , earthy <sup>d</sup>
3-Isobutyl-2-methoxypyrazine	7	38	14	56.10	6	59–97 <sup>a</sup>	0.002 <sup>a</sup>	Peasy <sup>a</sup> , earthy <sup>d</sup>
<b>Terpenes</b>								
Linalool	15	143	57	51.73	20	780–1310 <sup>c</sup>	0.17 <sup>a</sup>	Flowery <sup>a</sup> , floral, fruity (citrus) <sup>c</sup>
Geraniol	5	75	36	62.13	16	–	1.1 <sup>a</sup>	–
Limonene	29	169	80	44.07	43	1080–1320 <sup>c</sup>	4 <sup>a</sup>	–

CV (%) coefficient of variation, LSD ( $P < 0.05$ ) least significant difference at  $P$  value less than 0.05

<sup>a</sup> Summarized by Sunarharum et al. (2014)

<sup>b</sup> Summarized by Amanpour et al. (2015)

<sup>c</sup> Summarized by Piccino et al. (2014)

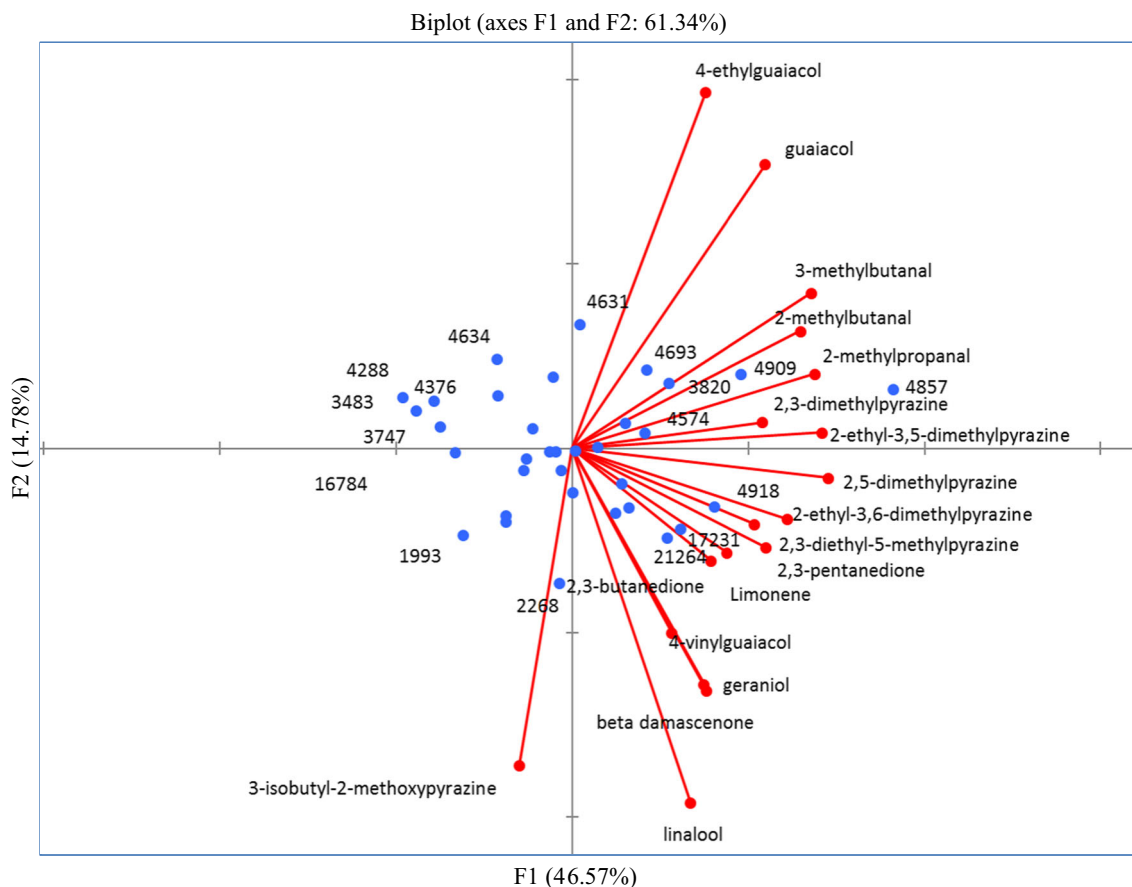
<sup>d</sup> Summarized by Belitz et al. (2009)

compounds (guaiacol, 4-ethylguaiacol), while accessions 4631, 4634 and 4693 were the opposite.

### Correlations among bean physical and biochemical traits

Spearman rank correlation analyses of the green bean morphology and non-volatiles among 232 accessions showed that there were strong correlations between bean physical parameters (Table 4). However, there was a small but significant positive correlation between caffeine and trigonelline ( $r = 0.141$ ,  $P < 0.05$ ) and between caffeine and 100 bean weight ( $r = 0.199$ ,  $P < 0.001$ ), while trigonelline had a slight negative correlation with 100 bean weight ( $r = -0.157$ ,  $P < 0.05$ ) (Table 4).

Among the 35 accessions selected for volatile compound analysis, there was no strong correlation between volatile profiles and either bean morphology or non-volatile levels, except for a strong correlation between bean roasting degree (or roasting colour) and 4-ethylguaiacol ( $r = -0.79$ ,  $P < 0.001$ ) or guaiacol ( $r = -0.75$ ,  $P < 0.001$ ) (i.e. higher 4-ethylguaiacol or guaiacol in darker roasted bean) or less strong correlation with 3-methylbutanal ( $r = -0.42$ ,  $P < 0.001$ ). Strong correlations existed between compounds belong to the same class (aldehydes, phenolic compounds, pyrazines) (up to  $r = 0.93$ ,  $P < 0.001$ ) or between compounds of different classes, such as aldehydes and ketones ( $r = 0.73$ ,  $P < 0.001$ ), aldehydes and phenolic compounds ( $r = 0.82$ ,  $P < 0.001$ ) and aldehydes and pyrazines ( $r = 0.72$ ,  $P < 0.001$ ) (Table 5).



**Fig. 2** Principle component analysis of 18 volatiles measured in roasted coffee bean for 35 arabica accessions using targeted analysis (red lines and dots are volatile compounds; blue dots are accession codes)

**Discussion**

**Population diversity and its relevance to quality breeding**

*Bean morphology*

No significant correlation between physical characteristics of green coffee bean and beverage quality has been reported (Kathurima et al. 2009). In this study, green bean physical characteristics and coffee quality of a large number of individuals (232 arabica coffee genotypes) were assessed, from which a confident result on the relationship between these

attributes was obtained. For the physical bean quality, the weight of 100 beans per accession was used in order to provide a representative measure of the bean size. Alternatively, the bean size can be measured directly on the sieve used in industry. Coffee with larger beans usually gets a good grade and fetches a higher price than coffee with smaller beans even though the former do not necessarily produce a more desirable roast or liquor (Belete 2014).

The large variation in bean size in the population studied based on visual assessment and weight of 100 beans (W100) (Table 1) indicated the potential for selection in breeding. In general, the population used in this study had a smaller

**Table 4** Spearman rank correlation between caffeine and trigonelline content (in green bean) and green bean physical characteristics (*n* = 232)

Variables	Caffeine	Trigonelline	W100	Density	Length	Width
Trigonelline	0.141*					
W100	0.199**	-0.157*				
Density	-0.106	0.000	-0.232***			
Length	0.077	0.041	0.567***	-0.195**		
Width	0.118	-0.106	0.564***	-0.217***	0.267***	
Thickness	0.160*	-0.056	0.598***	-0.222***	0.445***	0.382***

\*, \*\*, \*\*\*Significantly different from zero at the 0.05, 0.01 and 0.001 significance level, respectively



**Table 5** Spearman rank correlation between bean morphology, non-volatiles and volatiles of 35 samples

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)		
(2)	-0.12																								
(3)	0.13	-0.22																							
(4)	0.23	-0.41	0.25																						
(5)	0.07	0.44	-0.08	-0.08																					
(6)	-0.47	0.37	-0.26	-0.06	0.20																				
(7)	-0.11	-0.29	0.15	-0.03	-0.21	-0.31																			
(8)	0.10	-0.21	0.07	-0.04	-0.16	-0.37	<b>0.68</b>																		
(9)	0.12	-0.12	0.01	-0.22	-0.09	-0.28	<b>0.53</b>	<b>0.82</b>																	
(10)	-0.04	0.04	-0.06	-0.42	-0.07	-0.24	<b>0.50</b>	<b>0.75</b>	<b>0.93</b>																
(11)	-0.09	-0.03	0.01	-0.27	-0.14	-0.23	<b>0.54</b>	<b>0.73</b>	<b>0.92</b>	<b>0.92</b>															
(12)	-0.28	0.30	-0.28	<b>-0.75</b>	0.02	0.04	0.20	0.35	<b>0.55</b>	<b>0.69</b>	<b>0.57</b>														
(13)	-0.20	0.44	-0.39	<b>-0.79</b>	0.16	0.13	-0.09	0.05	0.31	<b>0.49</b>	0.35	<b>0.86</b>													
(14)	0.18	-0.29	0.36	0.01	-0.09	-0.18	0.32	0.37	0.27	0.25	0.22	0.19	0.13												
(15)	0.17	-0.30	0.41	0.08	-0.11	-0.24	0.40	<b>0.56</b>	<b>0.46</b>	0.38	0.36	0.11	-0.06	<b>0.51</b>											
(16)	-0.03	-0.09	0.14	-0.11	0.13	-0.14	<b>0.48</b>	<b>0.67</b>	<b>0.72</b>	<b>0.68</b>	<b>0.58</b>	<b>0.51</b>	0.25	0.38	<b>0.54</b>										
(17)	-0.01	-0.08	0.06	0.09	0.30	-0.04	0.41	<b>0.52</b>	<b>0.55</b>	<b>0.48</b>	0.43	0.34	0.07	0.30	0.38	<b>0.81</b>									
(18)	0.00	-0.08	0.02	0.16	0.35	0.00	0.26	0.34	0.38	0.31	0.25	0.26	0.03	0.18	0.21	<b>0.69</b>	<b>0.93</b>								
(19)	0.11	-0.11	0.23	-0.21	0.03	-0.06	0.13	0.27	<b>0.48</b>	0.38	0.37	0.42	0.28	0.34	0.45	<b>0.61</b>	0.41	0.38							
(20)	-0.08	0.00	-0.01	-0.15	0.25	-0.08	0.44	<b>0.56</b>	<b>0.66</b>	<b>0.64</b>	<b>0.56</b>	<b>0.57</b>	0.33	0.33	0.37	<b>0.87</b>	<b>0.94</b>	<b>0.86</b>	<b>0.52</b>						
(21)	0.27	<b>-0.50</b>	<b>0.50</b>	0.37	0.09	-0.24	0.10	0.05	-0.07	-0.18	-0.18	-0.30	-0.39	0.47	0.37	0.13	0.17	0.20	0.12	0.03					
(22)	-0.03	-0.35	0.44	0.12	-0.17	-0.13	0.44	<b>0.50</b>	0.32	0.27	0.23	0.06	-0.26	0.32	<b>0.62</b>	<b>0.50</b>	0.42	0.33	0.32	0.38	0.43				
(23)	-0.12	-0.24	0.25	-0.04	-0.26	-0.03	0.37	0.27	0.28	0.21	0.24	0.19	-0.03	0.07	0.46	0.38	0.21	0.15	0.37	0.26	0.29	<b>0.72</b>			
(24)	-0.12	-0.12	0.14	0.05	0.03	-0.22	0.33	0.46	0.40	0.38	0.44	0.29	0.04	0.16	0.22	<b>0.47</b>	0.36	0.33	0.21	0.42	0.10	0.43	0.32		

(1) Weight of 100 bean, (2) bean moisture loss during roasting, (3) roasted bean density, (4) bean roasting colour, (5) caffeine, (6) trigonelline, (7) 2,3-butanedione, (8) 2,3-pentanedione, (9) 2-methylpropanal, (10) 3-methylbutanal, (11) 2-methylbutanal, (12) guaiacol, (13) 4-ethylguaiacol, (14) 4-vinylguaiacol, (15) beta damascenone, (16) 2,5-dimethylpyrazine, (17) 2-ethyl-3,6-dimethylpyrazine, (18) 2,3-diethyl-5-methylpyrazine, (19) 2,3-dimethylpyrazine, (20) 2-ethyl-3,5-dimethylpyrazine, (21) 3-isobutyl-2-methoxypyrazine, (22) linalool, (23) geraniol, (24) limonene. Values in italics are different from 0 at the 0.001 significance level

average size compared to those in previous studies, but the range of W100 and bean length was larger than that reported in the literature. Previous studies showed that the weight of 100 beans of arabica hybrids and maternal lines ranged from 17.00 to 20.01 g (Bertrand et al. 2005). A more diverse population of 148 accessions from Ethiopia showed a mean 100 bean weight of two groups of 16.17–16.63 g with a minimum of 12.30 g and a maximum of 20.8 g (moisture content of 10%) (Montagnon and Bouharmont 1996). A larger variation in 100 bean weight was observed for 30 accessions at four locations in Ethiopia with average of 13.12 to 19.13 with minimum of 9.77 g and maximum of 21.82 g (at 11% moisture content) (Belete 2014). With a lower limit for W100 of 10.16 g and upper of 23.13 exceeding other studies at both ends of the spectrum, this study has captured a greater range of diversity in bean weight than previous, smaller investigations. Accessions with exceptionally large bean size identified in the current study could be a useful source for genetic improvement of arabica physical quality.

Bean length, width, length/width ratio and thickness reflect the bean shape and also contribute to the 100 bean weight. In a previous study involving 52 samples collected from Kenya, the average bean length, width and thickness were reported as 11.04, 8.27 and 5.13 mm, respectively (Ghosh and Gacanja 1970). Similarly, Olukunle and Akinnuli (2012) also reported the measurement for beans collected from Nigeria with

average length, width and thickness of 8.19, 6.11 and 4.60 mm, respectively. The current study involving a much larger population exhibited smaller mean size in each of these dimensions.

Density is a parameter that relates to bean weight as well as the roasting properties. The population studied had relatively low density compared to previous studies (Franca et al. 2005; Rodrigues et al. 2009). This may be the result of the higher moisture content of studied samples than that of the standard which relates to bean weight and volume.

Among the three groups of arabica, the hybrids had the largest average bean size and the most variable size (Table 1), which could be a result of crossbreeding from both parents that had large bean size or selection of hybrid population from a subset of the progenies of the cross which was selected based on large bean size. This phenomenon on the positive genetic gain for bean size had also been reported (Leroy et al. 2006; Walyaro 1983). The groups of cultivars and wild relatives had almost the same value for bean size of 15.35 and 15.57 g, while the wild groups included more accessions and would be expected to include more variation. However, among the three groups, the differences are not significant for weight of 100 beans, length and bulk density. For width and thickness which relate to bean size and ratio of L/W which relates to bean shape, the wild-type group was significantly different from cultivar and hybrid groups

suggesting its potential use in selection for bean size and shape in quality breeding. Furthermore, the mean L/W ratio in our study was larger than reported in the literature (Table 2), indicating that a more elongated bean shape is observed when a large diverse sample population is examined.

#### *Non-volatile compounds*

Although caffeine and trigonelline contents in coffee have been analysed in a number of reports, this is the first study using a relatively large arabica population of 232 genotypes with many wild accessions. Most other studies involved fewer accessions with different sample sizes and species and focused more on caffeine than on trigonelline. Previous studies showed caffeine content ranging from 0.62 to 1.82% (Anthony et al. 1993; Avelino et al. 2005; Ky et al. 2001b; Martin et al. 1998; Mazzafera and Carvalho 1992; Mehari et al. 2016; Taveira et al. 2014) which is very similar to the current study (0.82–1.76%). The study by Silvarolla et al. (2000) showed a substantial variation from 0.42 to 2.90% but included a very large population of interhybrid and intrahybrid progenies with 499 plants from 68 progenies (Kaffa region of Ethiopia) and 166 plants from 22 progenies (Illubabor region of Ethiopia). For trigonelline, the current study showed smaller variation (0.80–1.38%) compared to previous studies, especially that reported in the study by Mazzafera (1991) (1.52–2.90%). However, the current study gave results that were very close to those of the recent study by Mehari et al. (2016) using 99 arabica coffee samples from eight varieties from Ethiopia. The differences could be genetic or due to different methods of extraction and/or analytical instrument used in different studies. For caffeine, the significantly lower concentrations in the cultivars relative to the hybrid and wild-type groups indicated the result of selection for lower caffeine in the cultivar group.

The demand for decaffeinated coffee is increasing and now accounts for 10% of the total coffee consumed in the world. Selection and breeding for low caffeine content is thus a new target of the coffee industry. Accessions that had low caffeine content in this study may serve as a source of desirable genes for development of varieties with low caffeine content.

Among the 232 accessions studied, the 10 lowest in caffeine content were mostly from the variety group (6 accessions); only 2 were from the wild-type group and 2 from the hybrid group. Among the 10 highest in caffeine content, 2 accessions were hybrids followed by 6 accessions from wild types and 2 from the cultivar group. Seven out of the 10 accessions with the lowest trigonelline come from Ethiopia wild types. It is interesting to observe that there were fewer wild-type accessions found in the groups which were low caffeine and high trigonelline as expected. The accessions that had low caffeine content or high trigonelline content may serve as a source of desirable genes for development of variety

types with relatively low caffeine and high trigonelline content. This study has defined the variation in key quality attributes in the Arabica gene pool. It suggests approaches that could be used in coffee breeding to deliver high-quality coffee varieties. However, further replicated trials across environments would be needed to confirm the genetic value of those accessions.

#### *Volatile compounds*

Significant variation in volatile concentrations among the 35 diverse accessions indicates their potential in quality improvement. Among 18 compounds quantified, 5 had concentrations in the range previously reported and 8 were lower than that reported in the literature. However, these studies were limited in the number of accessions examined (Cheong et al. 2013; Czerny et al. 1999; Czerny and Grosch 2000; Piccino et al. 2014; Semmelroch and Grosch 1996; Semmelroch et al. 1995). It is interesting that most compounds belonging to the pyrazine class (2,5-dimethylpyrazine, 2-ethyl-3,6-dimethylpyrazine, 2,3-dimethylpyrazine and 2-ethyl-3,5-dimethylpyrazine) were much higher than in the literature (summarized by Sunarharum et al. 2014) even with the same roasting degree (light to medium), for example, 2,5-dimethylpyrazine of 53,312 ppb in this study vs 662 ppb or 2,3-dimethylpyrazine of 11,963 vs 119 ppb in another study (Toci and Farah 2014). A study investigating the influence of environment on coffee volatiles found that aldehydes and ketones appeared to be positively linked to elevated temperatures and high solar radiation (Bertrand et al. 2012). The coffee used in the current study was collected from an elevation of 616 m (above sea level) and average annual temperature of 22.2 °C. This may partly explain the reason for the lower concentration of these compounds compared to those in the literature and the high pyrazines. Low variation among replicates (or % CV) suggests the reliability of the volatile analysed and that these volatiles such as 4-vinylguaiacol, beta damascenone, a few volatiles of pyrazines (2-ethyl-3,6-dimethylpyrazine and 2,3-diethyl-5-methylpyrazine) and aldehydes (3-methylbutanal) can be used as criteria for breeding purposes.

The multivariate analysis (PCA) of the 18 volatiles showed that almost all volatile compounds of the same chemical class were clustered into groups as expected indicating the reliability of the method used. The population was distributed with contrasting volatile profile (along the PC1), either low or high in almost all volatiles suggesting that the selection of samples for low or high volatiles for breeding is possible. It is interesting to observe that samples with high volatile contributing most to the PC1 were from wild-type group such as 4857 (E-457, Ethiopia, wild), 4909 (E-540, Ethiopia, wild), 4918 (E-549, Ethiopia, wild) and 17231 (L334 et 45 c7, Cameroon, wild), while samples with low volatile were mainly from the

variety/cultivar group and hybrid group, for example, 4288 (Irgalen Kella Sidano, Belgian Congo, cultivar), 3747 (San Rafael, Costa Rica, variety), 4387 (Hibrido de Timor CRRC 1343/80, Portugal, hybrid) and 16784 (Sarchimor F3 IAC 1669/31-1 CIFIC H-361/971-10 Villasarchi X 832/2, Brazil, hybrid). This indicates that the previous selection of varieties for cultivation was mainly based on bean morphology and that wild-type accessions could be a valuable source for breeding. Similarly, accessions contributing most to PC2 which are higher in phenolic compounds (guaiacol—phenolic, burnt and 4-ethylguaiacol—spicy) and lower in linalool (terpenes—flowery aroma and 3-isobutyl-2-methoxypyrazine (pyrazines)—peasy such as 4631 (E-358, Ethiopia, wild), 4634 (E-361, Ethiopia, wild), 4693 (Limnu E-188, Ethiopia, wild) and 1993 (Goiaba colección 11 (552), Brazil, selection)) or accessions which were opposite in the concentration of these compounds such as 2268 (San Martin, Guatemala, cultivar) and 21264 (ET-08, Indonesia, wild) could be useful materials for genetic improvement of the corresponding compounds.

### Trait relationship and implication to quality breeding

The lack of strong correlation between physical bean properties and chemical compounds in this study agrees with other reports. Trugo et al. (1983) found no significant correlation between caffeine and trigonelline contents in 13 instant coffees. Similarly, Casal et al. (2000) analysed 20 roasted robusta coffees and 9 roasted arabica coffees and found that while there was a strong negative correlation between caffeine and trigonelline when all data from both arabica and robusta samples was analysed simultaneously, their correlation was not significant when coffee varieties were analysed separately. Kathurima et al. (2009) studied the relationship between beverage quality and green bean physical characteristics of 42 arabica coffee genotypes from Kenya and found a relatively better beverage quality in the beans with low 100 bean weight compared to that of larger beans, but the correlation was not significant. Similarly, Dessalegn et al. (2008a) indicated that beans with a low 100 bean weight had relatively more caffeine than heavy beans in 42 *C. arabica* genotypes from Ethiopia. This study also showed that correlations between caffeine content and green bean physical characteristics were not significant, while caffeine content had a significant negative correlation with all cup quality attributes of coffee such as acidity, body, flavour and overall standard of the liquor. However, these studies were implemented with only 42 genotypes.

The large variation observed for bean size of the germplasm collection (Table 1) provides a useful opportunity for improving genetic gain, as it is also a highly heritable trait and may have an effect on quality. Giomo et al. (2012) used 24 *C. arabica* genotypes from Brazil to evaluate physical

characteristics of the green beans and to describe the sensory profile. The results indicated that there were significant genotype effects both on coffee bean size and overall sensory quality, and the genotype  $\times$  environmental interaction was not significant. Thus, the green bean size could be considered an important criterion for coffee plant selection aiming to improve the green bean quality since the evaluations and comparisons have occurred in the same environmental conditions and in the same post-harvest processing procedures. In our study, however, the correlation between 100 bean weight and caffeine and trigonelline and between caffeine and trigonelline was not highly significant when investigated in a larger number of samples ( $n = 232$ ). This implied that it is difficult to select for low or high non-volatile compounds such as caffeine and trigonelline based on bean physical characteristics. However, it also indicates that the selection for this trait will not be affected by the selection of the others, or it is possible to select at the same time a variety that combines large bean size, low caffeine and high trigonelline. Development of molecular markers for each of these traits would enable them to be selected simultaneously in breeding. The investigation of a smaller subset of samples ( $n = 35$ ) based on two volatile extremes showed a strong negative correlation between the weight of 100 beans and trigonelline ( $r = -0.57$ ,  $P < 0.001$ ) indicating that trigonelline may be an aroma precursor linked to volatiles presented in the samples such as furans, pyrazines, pyrroles and pyridines (Franca et al. 2005). Based on the phenotypic data generated for this diverse arabica germplasm collection, further genotyping of the population will enable marker-trait association analysis for application in marker-assisted selection.

To our knowledge, this is the first report where correlations between coffee bean volatile compounds have been examined. A number of previous studies merely focused on the relationship between bean non-volatile and cupping quality which was controlled by the volatiles. Franca et al. (2005) and Farah et al. (2006b) found that caffeine had a positive correlation with high-quality beverages in arabica. Similarly, Figueiredo et al. (2013) reported that higher trigonelline was found in the best sensory score genotypes, and Barbosa et al. (2012) confirmed that both trigonelline and caffeine correlate with better sensory scores. However, Figueiredo et al. (2013) showed no correlation between caffeine content and beverage quality. Similarly, Avelino et al. (2005) studied arabica grown in two regions of Costa Rica and found that both caffeine and trigonelline were not well correlated with the sensory characteristics. The effect of environment to beverage quality score was also reported in several studies (Avelino et al. 2005; Barbosa et al. 2012; Rodrigues et al. 2009). The current study showed that both caffeine and trigonelline have weak positive and negative correlation with almost all volatiles even though trigonelline was considered as an important precursor of volatile compounds that link to coffee aroma and taste (Malta and

Chagas 2009). This would be hard to link to the sensory attributes as the complexity and the interaction of the volatiles in the aroma matrix exist.

He et al. (2015) found that darker roast correlated with aromatic compounds such as heptyl ether, 4-ethyl-2-methoxyphenol, 5-methylfuran-2-carbaldehyde and 1-(furan-2-ylmethyl)-1H-pyrrole, while Belitz et al. (2009) reported that 2-furfurylthiol and guaiacol increase with increasing degree of roasting. Mayer et al. (1999) found that pyrazines (2,3-diethyl-5-methylpyrazine, 2-ethenyl-3-ethyl-5-methylpyrazine, 3-isobutyl-2-methoxypyrazine), 4-vinylguaiacol, vanillin, 2-methyl-3-furanthiol and dimethyl trisulfide had almost no significant change with degree of roasting, while propanal, 2(5)-ethyl-4-hydroxy-5(2)-methyl-3(2H)-furanone, guaiacol, 4-ethylguaiacol, 2-furfurylthiol, 3-methyl-2-buten-1-thiol and methanethiol were affected by roasting. However, Toci's study (Toci and Farah 2014) showed that the concentration of almost all pyrazines decreased with roast degree. Holscher and Steinhart (1992) indicated that ketones (2,3-butanedione and 2,3-pentanedione) and aldehydes (2-methylpropanal, 3-methylbutanal, 2-methylbutanal) were decreased with degree of roasting. In the current study, we found that colour of roasted bean had a strong correlation with 4-ethylguaiacol and guaiacol and less strong correlation with 3-isobutyl-2-methoxypyrazine and 3-methylbutanal (Table 5) which is in agreement with Belitz et al. (2009) and Mayer et al. (1999) for guaiacol and most of pyrazines. However, it was different from the study of Holscher and Steinhart (1992) for ketones and aldehydes. The results one more time implied that the roasted bean colour could serve as a simple indicator for selecting certain favourable volatiles.

Strong and significant correlations between volatile compounds that belong to the same class (aldehydes, phenolic compounds, pyrazines) or between compounds of different classes (aldehydes and ketones, aldehydes and phenolic compounds, and aldehydes and pyrazines) suggest that the selection of compounds for analysis should be the indication for many more other volatiles. This could help reduce the labour, effort and cost for volatile analysis by focusing only on representing compounds that are easier and more stable to measure. For aldehydes, 3-methylbutanal can be selected for analysis as it is highly correlated with 2-methylbutanal ( $r = 0.91$ ,  $P < 0.001$ ) and 2-methylpropanal ( $r = 0.91$ ,  $P < 0.001$ ) and has lower variation among replicates. For phenolic compounds, 4-vinylguaiacol had a weak correlation with both guaiacol and 4-ethylguaiacol, while 4-vinylguaiacol had a highly reproducible result; it can be selected for analysis along with either guaiacol or 4-ethylguaiacol, which were strongly correlated. For pyrazines, among the six volatiles that were strongly correlated, 2-ethyl-3,6-dimethylpyrazine and 2,3-diethyl-5-methylpyrazine had the least variation among replicates (i.e. most reproducible) and thus could be used for analysis of this chemical class.

In summary, the observed high diversity in bean morphology, non-volatiles and volatiles in the worldwide arabica collection demonstrated its potential application as a valuable genetic resource to quality improvement in coffee breeding. Direct use of accessions with desirable traits observed in the study would require further multi-location genetic tests involving replications per accession. Biochemical analysis would still be needed to quantify both volatile and non-volatile compounds because bean morphology cannot be used as their predictor based on our correlation analysis. Most practically, the strong correlations existing within several volatile groups provide useful direction for targeted analyses focusing on reproducible and representing compounds so as to improve analytical accuracy and efficiency in coffee bean quality research and industry application.

**Acknowledgements** The authors would like to acknowledge the Tropical Agricultural Research and Higher Education Center (CATIE) for research germplasm, Peter Wolff—Wolff Coffee—for bean roasting and Wenny Sunarharum, Steve Fuller and Kent Fanning for advice and laboratory analysis.

**Data archiving statement** The list of germplasm used in the manuscript was provided in the [supplement data](#). Data on volatile compounds of 35 accessions used in the manuscript were also provided in the [supplement data](#).

**Author contributions** Tran, H. T.M. and Henry, R. J. conceptualized and outlined the manuscript and wrote the manuscript. Tran, H. T.M. and Vargas, C.A.C. collected coffee samples. Tran, H. T.M. carried out bean measurement and sample preparation for HPLC and GCMS analysis and performed data analyses. Smyth, H. conducted the volatile compound identification and quantification. Lee, L. S. and Furtado, A. contributed to the completion of the manuscript. All authors reviewed and approved the final manuscript.

**Compliance with ethical standards**

**Conflict of interest** The authors declare that they have no conflict of interest.

**Funding** This project is jointly supported by the Australian Research Council (ARC) and Green Cauldron.

## References

- Addinsoft (2007) XLSTAT, Analyse de données et statistique avec MS Excel. Addinsoft, NY
- Aerts R, Berecha G, Gijbels P, Hundera K, Glabeke SV, Vandepitte K, Muys B, Roldan-Ruiz I, Honnay O (2012) Genetic variation and risks of introgression in the wild *Coffea arabica* gene pool in South-Western Ethiopian montane rainforests. *Evol Appl*: 243–252. doi: [10.1111/j.1752-4571.2012.00285.x](https://doi.org/10.1111/j.1752-4571.2012.00285.x)
- Aga E, Bryngelsson T, Bekele E, Salomon B (2003) Genetic diversity of forest arabica coffee (*Coffea arabica* L.) in Ethiopia as revealed by random amplified polymorphic DNA (RAPD) analysis. *Hereditas* 138:36–46
- Agtron (2004) Agtron M-basic/II coffee roast analyser. Agtron Inc, Reno, p 10

- Amanpour, A., and Selli, S. (2015). Differentiation of volatile profiles and odor activity values of Turkish coffee and French press coffee. *Journal of Food Processing and Preservation*, 1-9. ISSN: 1745-4549
- Anthony F, Clifford MN, Noirot M (1993) Biochemical diversity in the genus *Coffea* L.: chlorogenic acids, caffeine and mozambioside contents. *Genet Resour Crop Evol* 40:61-70
- Anthony F, Bertrand B, Quiros O, Wilches A, Lashermes P, Berthaud J, Charrier A (2001) Genetic diversity of wild coffee (*Coffea arabica* L.) using molecular markers. *Euphytica* 118:53-65
- Anthony F, Astorga C, Avendaño J, Dulloo ME (2007a) Conservation of coffee genetic resources in the CATIE field genebank. In: Engelmann F, Dulloo ME, Astorga C, Dussert S, Anthony F (eds) *Conserving coffee genetic resources*. Biodiversity International, Rome, Italy
- Anthony F, Bertrand B, Astorga C, Lashermes P (2007b) Characterization and assessment of *Coffea arabica* L. genetic resources conserved in the CATIE field genebank. In: Engelmann F, Dulloo ME, Astorga C, Dussert S, Anthony F (eds) *Conserving coffee genetic resources*. Biodiversity International, Rome, Italy
- Avelino J, Barboza B, Araya JC, Fonseca C, Davrieux F, Guyot B, Cilas C (2005) Effects of slope exposure, altitude and yield on coffee quality in two altitude terroirs of Costa Rica, Orosi and Santa Mar'ia de Dota. *J Sci Food Agric* 85:1869-1876
- Barbosa JN, Borém FM, Cirillo MÂ, Malta MR, Alvarenga AA, Alves HMR (2012) Coffee quality and its interactions with environmental factors in Minas Gerais, Brazil. *J Agric Sci* 4:181-190
- Belay A (2011) Some biochemical compounds in coffee beans and methods developed for their analysis. *Int J Phys Sci* 6:6373-6378
- Belete Y (2014) Performance evaluations of hundred beans weights of indigenous Arabica coffee genotypes across different environments. *Sky J Agric Res* 3:120-127
- Belitz H-D, Grosch W, Schieberle P (2009) Food chemistry. In: Belitz H-D, Grosch W, Schieberle P (eds) *Chapter 21: Coffee, tea, cocoa*, 4th edn. Springer, Heidelberg
- Bertrand B, Etienne H, Cilas C, Charrier A, Baradat P (2005) *Coffea arabica* hybrid performance for yield, fertility and bean weight. *Euphytica* 141:255-262
- Bertrand B, Boulanger R, Dussert S, Ribeyre F, Berthiot L, Descroix F, Joet T (2012) Climatic factors directly impact the volatile organic compound fingerprint in green Arabica coffee bean as well as coffee beverage quality. *Food Chem* 135:2575-2583
- Bicchi C, Ruosi MR, Cagliero C, Cordero C, Liberto E, Rubiolo P, Sgorbini B (2011) Quantitative analysis of volatiles from solid matrices of vegetable origin by high concentration capacity headspace techniques: determination of furan in roasted coffee. *J Chromatogr A* 1218:753-762
- Buffo RA, Cardelli-Freire C (2004) Coffee flavour: an overview. *Flavour Fragr J* 19:99-104
- Campa C, Ballester JF, Doubeau S, Dussert S, Hamon S, Noirot M (2004) Trigonelline and sucrose diversity in wild *Coffea* species. *Food Chem* 88:39-43
- Campa C, Doubeau S, Dussert S, Hamon S, Noirot M (2005a) Diversity in bean caffeine content among wild *Coffea* species: evidence of a discontinuous distribution. *Food Chem* 91:633-637
- Casal S, Oliveira MB, Ferreira MA (1998) Development of an HPLC/diode-array detector method for simultaneous determination of trigonelline, nicotinic acid, and caffeine in coffee. *J Liq Chromatogr Relat Technol* 21:3187-3195
- Casal S, Oliveira MBPP, Alves MR, Ferreira MA (2000) Discriminate analysis of roasted coffee varieties for trigonelline, nicotinic acid, and caffeine content. *J Agric Food Chem* 48:3420-3424
- Charrier A, Eskes AB (2004) Botany and genetics of coffee. In: Wintgens JN (ed) *Coffee: Growing, Processing, Sustainable Production - A Guidebook for Growers, Processors, Traders, and Researchers*. WILEY-VCH Verlag GmbH & Co. KCA, Weinheim, Germany
- Cheong MW, Tong KH, Ong JJM, Liu SQ, Curran P, Yu B (2013) Volatile composition and antioxidant capacity of Arabica coffee. *Food Res Int* 51:388-396
- Czerny M, Grosch W (2000) Potent odorants of raw arabica coffee. Their changes during roasting. *J Agric Food Chem* 48:868-872
- Czerny M, Mayer F, Grosch W (1999) Sensory study on the character impact odorants of roasted arabica coffee. *J Agric Food Chem* 47: 695-699
- Davis AP, Tosh J, Ruch N, Fay MF (2011) Growing coffee: *Psilanthus (Rubiaceae)* subsumed on the basis of molecular and morphological data; implications for the size, morphology, distribution and evolutionary history of *Coffea*. *Bot J Linn Soc* 167:357-377
- De Maria CAB, Trugo LC, Moreira RFA (1995) Simultaneous determination of total chlorogenic acid, trigonelline and caffeine in green coffee samples by high performance gel filtration chromatography. *Food Chem* 52:447-449
- Dessalegn Y, Labuschagne MT, Osthoff G, Herselman L (2008a) Genetic diversity and correlation of bean caffeine content with cup quality and green bean physical characteristics in coffee (*Coffea arabica* L.) *J Sci Food Agric* 88:1726-1730
- Dessalegn Y, Herselman L, Labuschagne MT (2008b) AFLP analysis among Ethiopian arabica coffee genotypes. *Afr J Biotechnol* 7: 3193-3199
- Farah A (2009) Coffee as a speciality and functional beverage. In: Paquin P (ed) *Functional and speciality beverage technology*. Woodhead Publishing in Food Science, Technology and Nutrition, pp 370-395. Woodhead Publishing Limited, Cambridge
- Farah A, Monteiro MC, Calado V, Franca AS, Trugo LC (2006b) Correlation between cup quality and chemical attributes of Brazilian coffee. *Food Chem* 98:373-380
- Figueiredo LP, Borém FM, Cirillo MÂ, Ribeiro FC, Giomo GS, Salva TDJG (2013) The potential for high quality bourbon coffees from different environments. *J Agric Sci* 5:87-98
- Flament I (2002) *Coffee flavour chemistry*, Chichester, UK
- Franca AS, Mendonca JCF, Oliveira SD (2005) Composition of green and roasted coffees of different cup qualities. *Food Sci Technol* 38: 709-715
- Ghosh BN, Gacanja W (1970) A study of the shape and size of wet parchment coffee beans. *J Agric Eng Res* 15(2):91-99
- Giomo G, Borem F, Saath R, Mistro J, Figueiredo L, Ribeiro F, Pereira S, Bernardi M (2012) Evaluation of green bean physical characteristics and beverage quality of arabica coffee varieties in Brazil. 24th International Conference on Coffee Science 12th -16th November 2012, San José (Costa Rica)
- He Y, Hu R, Zhang H, Wen N, Cai T, Peng J, Xu Y (2015) Characteristic aroma detection of coffee at different roasting degree based on electronic nose. *Trans Chin Soc Agric Eng* 31:247-255
- Holscher W, Steinhart H (1992) Investigation of roasted coffee freshness with an improved headspace technique. *Eur Food Res Technol* 195: 33-38
- IPGRI (1996) *Descriptors for coffee (Coffea spp. and Psilanthus spp.)*. International Plant Genetic Resources Institute, Rome, pp 28-29
- Jeszka-Skowron M, Zgoła-Grzes'kowiak A, Grzes'kowiak T (2015) Analytical methods applied for the characterization and the determination of bioactive compounds in coffee. *Eur Food Res Technol* 240:19-31
- Kathurima C, Gichimu B, Kenji G, Muhoho S, Boulanger R (2009) Evaluation of beverage quality and green bean physical characteristics of selected Arabica coffee genotypes in Kenya. *Afr J Food Sci* 3: 365-371
- Ky C-L, Guyot B, Louam J, Hamon S, Noirot M (2001a) Trigonelline inheritance in the interspecific *Coffea pseudozanguebariae* × *C. liberica* var. *dewevrei* cross. *Theor Appl Genet* 102:630-634
- Ky C-L, Louam J, Dussert S, Guyot B, Hamon H, Noirot M (2001b) Caffeine, trigonelline, chlorogenic acids and sucrose diversity in

- wild *Coffea arabica* L. and *C. canephora* P. accessions. Food Chem 75:223–230
- Leroy T, Ribeyre F, Bertrand B, Chametant P, Dufour M, Montagnon C, Marraccini P, Pot D (2006) Genetics of coffee quality. Braz J Plant Physiol 18:229–242
- Malta MR, Chagas SJR (2009) Avaliação de compostos não-voláteis em diferentes cultivares de cafeeiro produzidas na região Sul de Minas Gerais. Acta Sci Agron 31:57–61
- Martens H, Karstang T, Næs T (1987) Improved selectivity in spectroscopy by multivariate calibration. J Chemom 1:201–219
- Martín MJ, Pablos F, González AG (1998) Discrimination between arabica and robusta green coffee varieties according to their chemical composition. Talanta 46:1259–1264
- Mayer F, Czerny M, Grosch W (1999) Influence of provenance and roast degree on the composition of potent odorants in Arabica coffees. Eur Food Res Technol 209:242–250
- Mazzafera P (1991) Trigonelline in coffee. Phytochemistry 30:2309–2310
- Mazzafera P, Carvalho A (1992) Breeding for low seed caffeine content of coffee (*Coffea* L.) by interspecific hybridization. Euphytica 59: 55–60
- Mehari B, Redi-Abshiro M, Chandravanshi BS, Atlabachew M, Combrinck S, McCrindle R (2016) Simultaneous determination of alkaloids in green coffee beans from Ethiopia: chemometric evaluation of geographical origin. Food Anal Methods. 9:1627–1637
- Montagnon C, Bouharmont P (1996) Multivariate analysis of phenotypic diversity of *Coffea arabica*. Genet Resour Crop Evol 43:221–227
- Nuhu AA (2014) Review article: bioactive micronutrients in coffee: recent analytical approaches for characterization and quantification. ISRN Nutr 2014:1–13
- Oestreich-Janzen S (2010) Chemistry of coffee. In: M L, Liu H-W (eds) Comprehensive Natural Products II: Chemistry & Biology, pp. 1085–1117 CAFE A GmbH, Hamburg
- Olukunle OJ, Akinnuli BO (2012) Investigating some engineering properties of coffee seeds and beans. J Emerg Trends Eng Appl Sci 3(5): 743–747
- Payne RW, Harding SA, Murray DA, Soutar DM, Baird DB, Glaser AI, Channing IC, Welham SJ, Gilmour AR, Thompson R, Webster R (2008) GENSTAT release 11 reference manual. Parts 1, 2 and 3. VSN International, Hemel Hempstead
- Piccino S, Boulanger R, Descroix F, Sing ASC (2014) Aromatic composition and potent odorants of the “specialty coffee” brew “Bourbon Pointu” correlated to its three trade classifications. Food Res Int 61: 264–271
- Pickard S, Becker I, Merz K-H, Richling E (2013) Determination of the alkyprazine composition of coffee using stable isotope dilution-gas chromatography-mass spectrometry (SIDA-GC-MS). J Agric Food Chem 61:6274–6281
- Rodrigues CI, Maia R, Miranda M, Ribeirinho M, Nogueira JMF, Maguas C (2009) Stable isotope analysis for green coffee bean: a possible method for geographic origin discrimination. J Food Compos Anal 22:463–471
- Semmelroch P, Grosch W (1996) Studies on character impact odorants of coffee brews. J Agric Food Chem 44:537–543
- Semmelroch P, Laskawy G, Blank I, Grosch W (1995) Determination of potent odorants in roasted coffee by stable isotope dilution assays. Flavour Fragr J 10:1–7
- Silvarolla MB, Mazzafera P, Lima MMAD (2000) Caffeine content of Ethiopian *Coffea arabica* beans. Genet Mol Biol 23:213–215
- Sunarharum W (2016) The compositional basis of coffee flavour. Queensland Alliance for Agriculture and Food Innovation. PhD thesis at The University of Queensland
- Sunarharum WB, Williams DJ, Smyth HE (2014) Review: complexity of coffee flavor: a compositional and sensory perspective. Food Res Int 62:315–325
- Taveira JHDS, Borém FM, Figueiredo LP, Reis N, Franca AS, Harding SA, Tsai C-J (2014) Potential markers of coffee genotypes grown in different Brazilian regions: a metabolomics approach. Food Res Int 61:75–82
- Teressa A, Crouzillat D, Petiard V, Brouhan P (2010) Genetic diversity of Arabica coffee (*Coffea arabica* L.) collections. EJAST 1:63–79
- Toci AT, Farah A (2014) Volatile fingerprint of Brazilian defective coffee seeds: corroboration of potential marker compounds and identification of new low quality indicators. Food Chem 253:298–314
- Tran HT (2005) Genetic variation in cultivated coffee (*Coffea arabica* L.) accessions in northern New South Wales, Australia. Southern Cross University, Lismore, NSW, Australia
- Trugo LC (1984) HPLC in coffee analysis. University of Reading, England
- Trugo LC, Macrae R (1989) Application of high performance liquid chromatography to the analysis of some non-volatile coffee compounds. Archivos Latinoamericanos de Nutricion 39:96–107
- Trugo LC, Macrae R, Dick J (1983) Determination of purine alkaloids and trigonelline in instant coffee and other beverages using high performance liquid chromatography. J Sci Food Agric 34:300–306
- Walyaro D (1983) Considerations in breeding for improved yield and quality in arabica coffee (*Coffea arabica* L.). Wageningen, The Netherlands
- WCR (2014) Assessment of genetic diversity in *Coffea arabica*. World Coffee Research 2014 Annual Report