Developing Single Nucleotide Polymorphism (SNP) Markers for the Identification of Coffee Germplasm

Lin Zhou[1,2] · Fernando E. Vega[1] · Huawei Tan[1,2] · Aixa E. Ramírez Lluch[3] · Lyndel W. Meinhardt[1] · Wanping Fang[1,2] · Sue Mischke[1] · Brian Irish[4] · Dapeng Zhang[1]

Received: 30 September 2015 / Accepted: 8 April 2016 / Published online: 20 April 2016

Abstract Coffee is one of the most widely consumed beverages and represents a multibillion-dollar global industry. Accurate identification of coffee cultivars is essential for efficient management, exchange, and use of coffee genetic resources. To date, a universal platform that can allow data comparison across different laboratories and genotyping platforms has not been developed by the coffee research community. Using expressed sequence tags (EST) of *Coffea arabica, C. canephora* and *C. racemosa* from public databases, we developed 7538 single nucleotide polymorphism (SNP) markers and selected 180 for validation using 25 *C. arabica* and C. *canephora* accessions from Puerto Rico. Based on the validation result, we designated a panel of 55 SNP markers that are polymorphic across the two species. The average minor allele frequency and information index of this SNP panel are 0.281 and 0.690, respectively. This panel enabled the differentiation of all tested accessions of *C. canephora*, which accounts for 79.2 % of the total polymorphism in the samples. Only 21.8 % of the polymorphic SNPs were detected in the 12 *C. arabica* cultivars, which, nonetheless, were able to unambiguously differentiate the 12 Arabica cultivars into ten unique genotypes, including two synonymous groups. Several local Puerto Rican cultivars with partial Timor pedigree, including Limaní, Frontón, and TARS 18087, showed substantial genetic difference from the other common Arabica cultivars, such as Catuai, Borbón, and Mundo Nuevo. This coffee SNP panel provides robust and universally comparable DNA fingerprints, thus can serve as a genotyping tool to assist coffee germplasm management, propagation of planting material, and coffee cultivar authentication.

Keywords Fluidigm · *Coffea arabica* · *Coffea canephora* · Molecular markers · Tropical agriculture

Abbreviations

cDNA	Complementary DNA
DNA	Deoxyribonucleic acid
EST	Expressed sequence tag
PCR	Polymerase chain reaction
SNP	Single nucleotide polymorphism
SSR	Simple sequence repeat
USDA	United States Department of Agriculture,
ARS	Agricultural Research Service

Introduction

The genus *Coffea* (Rubiaceae) comprises 124 species (Davis et al. 2006; Davis 2010, 2011), with only two species comprising the bulk of commercially traded coffee: *Coffea*

Communicated by: Alan Andrade

Electronic supplementary material The online version of this article (doi:10.1007/s12042-016-9167-2) contains supplementary material, which is available to authorized users.

✉ Dapeng Zhang
Dapeng.Zhang@ars.usda.gov

[1] Sustainable Perennial Crops Laboratory, United States Department of Agriculture, Agricultural Research Service, USDA ARS BARC-W, 10300 Baltimore Avenue, Bldg. 001, Rm. 223, Beltsville, MD 20705, USA

[2] College of Horticulture, Nanjing Agricultural University, Nanjing 210095, Jiangsu, China

[3] Puerto Rico Department of Agriculture, P. O. Box 10163, Santurce, PR 00908, USA

[4] Tropical Agriculture Research Station, United States Department of Agriculture, Agricultural Research Service, Mayagüez, PR 00682, USA

arabica L. and *C. canephora* Pierre ex. A. Froehner (also known as Robusta). Coffee is grown in over 10 million hectares in ca. 80 tropical and subtropical countries (FAOSTAT 2014) and is especially important for the livelihoods of approximately 20 million coffee-farming families in Asia, Africa and Latin America (Osorio 2002; Gole et al. 2002; Lewin et al. 2004). The yearly value of the entire coffee industry has been estimated at US$173 billion (International Coffee Organization 2014).

Coffee germplasm is the ultimate resource for breeding new cultivars with improved agronomic traits and quality attributes. It is estimated that over 21,000 germplasm accessions, representing 70 *Coffea* species, are maintained in numerous coffee germplasm collections around the world (Anthony et al. 2001; 2007; Vega et al. 2008). Like many other tropical perennial crops, coffee germplasm collections are usually maintained as living trees in the field due to the unorthodox nature of the seeds. It is not unusual for coffee germplasm collections to have limited information on their correct identity. Consequently, misidentifications, redundancy, and other errors affect the genebank administration and operation.

Accurate genotype identification is essential to improve the efficiency of coffee germplasm management and use of germplasm in breeding. Different molecular markers, especially microsatellite markers, have been developed and applied to identify coffee genotypes and analyze genetic diversity (Moncada and McCouch 2004; Masumbuko and Bryngelsson 2006; Cubry et al. 2008; López-Gartner et al. 2009; Tshilenge et al. 2009; Missio et al. 2010; Vieira et al. 2010; Geleta et al. 2012; Razafinarivo et al. 2013; Leroy et al. 2014).

While these markers have significantly improved the management of coffee germplasm, resolving genotyping results from different laboratories has not been straightforward. It is difficult to standardize data generated on different genotyping platforms, and comparison of data is further complicated because the same alleles may be binned differently. Even on the same platform, analysis can be complicated by common PCR artifacts such as stutter due to slipped strand mispairing, which may lead to incorrect identification of an allele, and by diminished amplification of longer repeats, which can lead to scoring a heterozygote as homozygous and other inaccuracies that are not identical to the true genotype (Zhang et al. 2006).

Single nucleotide polymorphisms (SNPs) have clear advantage over previously used molecular markers in terms of their applications in germplasm management. These applications include, but are not limited to, identification of mislabeled accessions, parentage and sibship analysis for quality control in breeding and seeds programs, and genetic authentication and traceability to support the production of high-value cultivars for premium markets. Compared to simple sequence repeat (SSR) markers, SNP analysis can be done without requiring DNA separation by size and can, therefore, be automated in high-throughput assay formats. The genotyping profiles of SNPs can be compared across different laboratories and genotyping platforms. These advantages have resulted in the increasing use of SNPs as the markers of choice for accurate genotype identification in tropical perennial crops, as recently demonstrated in cacao (*Theobroma cacao*; Ji et al. 2013), pummelo (*Citrus maxima*; Wu et al. 2014), tea (*Camellia sinensis*; Fang et al. 2014), longan (*Dimocarpus longan*; Wang et al. 2015), and litchi (*Litchi chinensis*; Liu et al. 2015). Nonetheless, this most powerful tool for germplasm management has not been directly applied to coffee germplasm identification.

Ample genomic resources have been developed for coffee (de Kochko et al. 2010; Krishnan and Ranker 2012; Hamon et al. 2015), and the genome of *C. canephora* has been sequenced (Denoeud et al. 2014). These resources provide opportunities for mining new markers that can be used for coffee germplasm management and breeding. Several studies involving SNPs discovery in coffee through mining of expressed sequence tags (ESTs) or transcriptome data have been published (de Kochko et al. 2010; Vidal et al. 2010; Combes et al. 2013; Yuyama et al. 2016). Recently, Genotyping by Sequencing (GBS) was applied to generate SNP markers for QTL mapping of agronomic traits in coffee (Moncada et al. 2016). However, so far no effort has been devoted to validate SNPs for coffee genotype identification. Our objective was to develop a set of SNP markers to assist coffee germplasm identification. The results reported herein represent a validation study of SNPs in coffee germplasm identification, demonstrating the utility of existing EST sequences as an approach for rapid development of a high quality genotyping tool. These SNP markers, as well as the genotyping method, will be particularly useful for germplasm management, including identification of mislabeled cultivar/accessions, parentage/sibship analysis, quality control in seed propagation, and intellectual property rights in cultivar protection.

Materials and Methods

Mining of Putative SNPs from Coffee EST Database

EST sequences of *C. arabica, C. canephora* and *C. racemosa* were obtained from the NCBI GenBank EST database (http://www.ncbi.nlm.nih.gov/). These EST sequences were developed from a large number of cDNA libraries of the three *Coffea* species, representing specific stages of cells and plant development (Vieira et al. 2006; Mondego et al. 2011; Moore and Ming 2008). The downloaded sequences were merged into a single dataset for data mining. The low quality sequences and poly-A tail segments of EST sequences in the dataset were removed using the program EST-TRIMMER (http://pgrc.ipk-gatersleben.de/misa/download/est_trimmer.pl). The cleaned sequences were then assembled to contigs using CAP3 Sequence Assembly Program (Huang and Madan 1999) with a 98 % sequence similarity threshold. Putative EST-SNPs were detected using the QualitySNPng program (Nijveen et al. 2013); http://www.bioinformatics.nl/QualitySNPng/). Only clusters that included at least six EST sequences, with a confidence score over 5, were accepted. In order to meet the requirements and constraints for primer design, all candidates for SNP markers with less than 60 nucleotides between two neighboring SNPs, and with flanking sequences less than 100 nucleotides long, were removed. The annotation of contigs was done by performing BLAST searches against the NCBI NR database and the *C. canephora* genome. A subset of the identified SNP sequences was then chosen for design and manufacture of primers to assay for SNPs in the coffee plant.

Validation of Putative SNPs

To evaluate the putative SNP markers for suitability of varietal identification, we used the EP1™ nanofluidic genotyping system (Fluidigm Corporation, San Francisco, CA) and validated the SNPs for 12 *C. arabica* and 13 *C. canephora* cultivars (Table 1). All samples were cultivated coffee accessions provided by the Puerto Rico Department of Agriculture, with the exception of TARS 18087, which was provided by the USDA, ARS, Tropical Agriculture Research Station in Mayagüez, Puerto Rico. The *C. arabica* cultivars were introduced to Puerto Rico as seeds from various countries, including Martinique, El Salvador, Indonesia and Portugal, whereas the *C. canephora* germplasm was introduced via a tree given by the French government and from seeds originating in Java, Indonesia (McClelland 1924; Várzea et al. 2009; Monroig Inglés n. d.).

DNA was extracted from dried coffee leaves with the DNeasy® Plant Mini kit (Qiagen Inc., Valencia, CA), which is based on the use of silica as an affinity matrix. The dry leaf tissue was placed in a 2 ml microcentrifuge tube with one ¼-inch ceramic sphere and 0.15 g garnet matrix (Lysing Matrix A; MP Biomedicals, Solon, OH). The leaf samples were disrupted by high-speed shaking in a TissueLyser II (Qiagen Inc., Valencia, CA) at 30 Hz for 1 min. Lysis solution (DNeasy® kit buffer AP1 containing 25 mg/ml polyvinylpolypyrrolidone), along with RNase A, was added to the powdered leaf samples and the mixture was incubated at 65 °C, as specified in the instructions. The remainder of the extraction method followed manufacturer's instructions. DNA was eluted from the silica column with two washes of 50 μL buffer AE, which were pooled, resulting in 100 μL DNA solution. Using a NanoDrop spectrophotometer (Thermo Scientific, Wilmington, DE), DNA concentration was determined by absorbance at 260 nm. DNA purity was estimated by the 260/280 ratio and the 260/230 ratio.

Table 1 List of 25 accessions of *C. arabica* and *C. canephora* from Puerto Rico used in the experiment

Sample code	Sample name	Species
Cof15	Limaní	*C. arabica* x Timor hybrid
Cof17	Frontón	*C. arabica* x Timor hybrid
Cof16	Puerto Rico	*C. arabica*
Cof19	Catuai	*C. arabica*
Cof18	Caturra	*C. arabica*
Cof21	Borbón	*C. arabica*
Cof22	Pacas	*C. arabica*
Cof23	Sumatra	*C. arabica*
Cof26	Mundo Nuevo	*C. arabica*
Cof24	Enrea	*C. arabica*
Cof25	Harrar-Rojo	*C. arabica*
Cof71	TARS 18087	*C. arabica*
Cof20	Robusta EEA	*C. canephora*
Cof58	A5-00726	*C. canephora*
Cof59	A5-00738	*C. canephora*
Cof61	A5-00742	*C. canephora*
Cof62	A5-00797	*C. canephora*
Cof63	A5-00906	*C. canephora*
Cof64	B4-00139	*C. canephora*
Cof65	B4-00140	*C. canephora*
Cof66	B4-00262	*C. canephora*
Cof67	B4-00265	*C. canephora*
Cof68	B4-00268	*C. canephora*
Cof69	Robusta-#69	*C. canephora*
Cof70	Robusta-#70	*C. canephora*

One hundred and eighty putative SNP sequences were submitted to the Assay Design Group at Fluidigm Corporation (San Francisco, CA) for design and manufacture of primers for an SNPtype™ genotyping panel. The assays were based on competitive allele-specific PCR and enable bi-allelic scoring of SNPs at specific loci (KBioscience Ltd., Hoddesdon, UK). The Fluidigm SNPtype™ Genotyping Reagent Kit was used according to the manufacturer's instructions (Fluidigm 2011). Using these primers, the isolated DNAs were subjected to Specific Target Amplification (STA) in order to enrich the SNP sequences of interest. Genotyping was performed on a nanofluidic 96.96 Dynamic Array™ IFC (Integrated Fluidic Circuit; Fluidigm Corporation, San Francisco, CA). This chip automatically assembles PCR reactions, enabling simultaneous testing of up to 96 samples with 96 SNP markers. End-point fluorescent images of the 96.96 IFC were acquired on an EP1™ imager and the data was analyzed with Fluidigm

Genotyping Analysis Software (Fluidigm Corporation, San Francisco, CA).

Data Analysis

Key descriptive statistics for measuring the informativeness of the SNP markers were calculated, including minor allele frequency (MAF), observed heterozygosity, expected heterozygosity, information index, and inbreeding coefficient. The program GenAlEx 6.5 was used for computation (Peakal and Smouse 2012). For genotype identification, pairwise multilocus matching was applied among individual samples, using the same program. DNA samples that were fully matched at the genotyped SNP loci were declared same genotype (or clones).

Distance-based multivariate analysis was used to assess the relationship among the individual farmer cultivars, as well as their relationship with reference clones from international genebanks. Pairwise genetic distances were computed using the DISTANCE procedure implemented in GenAlEx 6.5 (Peakall and Smouse 2012). The same program was then used to perform Principal Coordinates Analysis (PCoA), based on the pairwise distance matrix. Both distance and covariance were standardized.

As a complement to PCoA, a cluster analysis using the neighbor-joining algorithm was used to further examine the genetic relationship among accessions. Kinship coefficient was chosen as a genetic distance measurement of shared ancestry among the individual accessions. The computation was executed using Microsatellite Analyser (MSA; Dieringer and Schlötterer 2003). A dendrogram was generated from the resulting distance matrix using the neighbor-joining algorithm (Saitou and Nei 1987) available in PHYLIP (Felsenstein 1989). The unrooted tree was visualized using NJ Plot (Perrière and Gouy 1996).

Results

In Silico Analysis and SNP Discovery

A total of 174275, 69066 and 10838 EST sequences of *C. arabica*, *C. canephora*, and *C. racemosa,* respectively, were downloaded from NCBI (Jan 23, 2015). After adapter removal, trimming and quality control, 225763 high quality sequences were selected. The CAP3 program (Huang and Madan 1999) was used to assembly sequences into 25115 contigs and 59036 singlets, with an average size of 6.64 sequences per contig, among which putative SNPs were detected in 2989 contigs using the QualitySNP program (Tang et al. 2006). Each of these selected clusters included a minimum of six EST sequences, whereas the minimum redundancy threshold required by QualitySNP is three, and the minimal confidence score is five. In total, we obtained 7538 putative EST-SNPs, of which 4706 were transition mutations, including 2463 C/T and 2243 A/G. There were 2473 transversions, including 661 A/T, 545 A/C, 560 T/G, and 707 C/G, along with 333 indels, 21 tri-allelic polymorphisms and five tetra-allelic polymorphisms. To select high quality SNPs for validation, candidate SNP sites with about 60 bp before and after the site were filtered. Germplasm diversity of *C. arabica* is narrower than other *Coffea* species (Anthony et al. 2007), so we specially selected more putative SNPs from the ESTs of *C. arabica* than from *C. canephora and C. racemosa.* A total of 180 SNPs were used for primer designing. Detailed information of these 180 putative SNPs is presented in Supplemental Table 1. The designed 180 primer sequences are presented in Supplemental Table 2.

Descriptive Statistics for Validated SNP Markers

Of the chosen 180 SNP markers, 156 (87 %) were successful in genotyping. The failures of the remaining 24 SNPs were likely due to sequence complexities or the presence of polymorphisms within the flanking sequences. However, among the successful 156 SNPs, 45 were monomorphic across the 25 coffee samples (i.e., only one SNP variant was identified in all individuals). In addition, there were 56 markers that did not show intra-specific polymorphism in either *C. arabica* or in *C. canephora*, but they showed polymorphism between these two species. These interspecific polymorphic markers are potentially useful when a larger number of coffee samples are screened. However, they were not included in the subsequent data analysis for the present study. The summary of the validation result is presented in Table 2.

A total of 55 polymorphic SNPs, which were reliably scored across the validation panel, were retained for further analysis. Of the 55 polymorphic SNPs, only 12 are polymorphic in *C. arabica*. The flanking sequences and SNPs of the 55 coffee SNPs are listed in Table 3. The minor allele frequencies of these SNPs ranged from 0.04 to 0.48 with an average of 0.281. The mean information index was 0.536, ranging from 0.168 to 0.692. The observed heterozygosity ranged from 0.040 to 0.920 with an average of 0.351, whereas the mean expected heterozygosity was 0.362 ranging from 0.077 to 0.499 (Table 4).

Genotype Identification

Among the *C. canephora* accessions, every accession has unique SNP profiles as shown by the result of multilocus matching, which is in agreement with the out-crossing nature of *C. canephora*. Among the *C. arabica* cultivars, two synonymous groups were detected by multilocus matching, where the same SNP genotypes were found in samples with different cultivar names (Table 5). The synonymous group #1 contains

Table 2 Validation result of 180 SNPs using Fluidigm's EP1™ nanofluidic genotyping system (Fluidigm Corporation, San Francisco, CA) in 12 *C. arabica* and 13 *C. canephora* accessions

Marker categories	# SNPs	%
No amplification	24	13.3
Inter-specific monomorphic	45	25.0
Inter-specific polymorphic but intra-specific monomorphic	56	31.1
Inter- and intra-specific polymorphic	55	30.6
Total	180	100

Catuai and Caturra, whereas the synonymous group #2 includes Borbón and Sumatra. Each pair of cultivars is identical as assessed by the 55 SNP markers. However, there are possible differences between each pair of cultivars resulting from somaclonal mutations, which was undetectable by the 55 SNP markers. The total molecular variance among the 13 *C. canephora* accessions is 198.8, whereas the total molecular variance among the 12 *C. arabica* accessions is 26.2. There is a large genetic difference between the two species (Fst = 0.247; $P < 0.001$) and the Robusta group showed much higher diversity than the Arabica group.

Genetic Relationship Among Tested Cultivars/Accessions

The genetic relationships among the tested cultivars/accessions are illustrated using PCoA (Fig. 1a). The plane of the first three PCoA axes accounted for 72 % of the total variation (first axis = 57.5 %, second = 8.4 %, and third = 6.1 %). In the plane of coordinate 1 vs. 2, all 12 cultivars of *C. arabica* formed a small homogenous group, whereas the 13 *C. canephora* accessions were clearly separated from the *C. arabica* group and scattered in a much wider area. When the *C. arabica* group was analyzed alone, substantial variation was revealed by PCoA (Fig. 1b). The plane of the first three PCoA axes accounted for 65.5 % of the total variation (first axis = 28.9 %, second = 22.1 %, and third = 14.5 %). Three Puerto Rican local cultivars, including Limaní, Frontón and TARS 18087 showed clear genetic differences from the founder Arabica cultivars such as Borbón, Puerto Rico, Catuai, and Caturra.

The unrooted neighbor-joining tree grouped the 25 accessions into two main clusters, representing Robusta and Arabica, respectively (Fig. 2). Within the cluster of Arabica cultivars, the genetic relationships are fully compatible with those shown for PCoA (Fig. 1b), where local Arabica cultivars, including Limaní, Frontón and TARS 18087, showed unique genetic profiles that differed substantially from the rest of the Arabica cultivars. Within the Robusta clusters, there is heterogeneity that divided the 13 accessions into two subgroups. The first subgroup includes A5 00738, B4 00140, B4 00268 and A5 00797, whereas the rest of the accessions fall into the second cluster, indicating there are different genetic origins for the Robusta coffee in Puerto Rico.

Discussion

Success Rate of Mining SNPs from ESTs

A number of genomic resources have been developed in coffee (Denoeud et al. 2014; Dereeper et al. 2015), but their application to assist germplasm management has been limited, and advanced molecular tools to support germplasm management are not available. Developing SNP markers from EST sequences has been considered an efficient strategy for non-model species. In the present study, we designed 7538 candidate SNPs through direct mining of coffee EST sequences stored in public databases, which were generated from various tissue and organs at different developmental stages. We then tested 180 candidate SNPs by genotyping a panel of 25 accessions of Arabica and Robusta coffee germplasm. We obtained a success rate of 31.1 % for marker validation, which is much lower compared with our recent studies that used the same approaches on other tropical perennial crops, such as tea (58 %; Fang et al. 2014) and pineapple (60 %; Zhou et al. unpublished data). This low success rate could be attributed mainly to the fact that the sample panel for validating the selected SNPs includes only 25 accessions/cultivars from Puerto Rico, which may not have enough diversity to show polymorphism.

Indeed, *C. arabica* originated from east Africa and only a small fraction of coffee genetic diversity was introduced into Latin America (Vega 2008). Moreover, it is estimated that the genetic variation in *C. arabica* is generally only about one tenth of that in *C. canephora* (Lashermes et al. 2000). The amount of genetic diversity in *C. arabica* cultivars grown in Latin America is even smaller, because these cultivars were mostly mutants derived from a single cultivar, or intra-cultivar selections from that cultivar ('Tipica'; Anthony et al. 2007). The success rate may increase as the size of the tested sample panel increases and/or when additional accessions from Africa (especially Ethiopia) are included. For this reason, we present in the supplementary table those SNPs that did not show intra-specific polymorphism, in case other researchers in the coffee community would like to test their usefulness in the future.

The present results also show that some SNP markers are species specific, and that SNPs were polymorphic in one species but monomorphic in the other species. The *C. arabica* specific SNPs can be explained by contribution from the *C. arabica* subgenome *C. eugenioides*. The *C. canephora* specific SNPs were likely due to the limited sampling of the *C. canephora* subgenome, because only a small fraction of the *C. canephora* genetic diversity was incorporated in cultivated *C. arabica*. When the 180 SNPs used in present assay were

Table 3 Flanking sequences and SNPs of the 55 polymorphic markers in 25 *C. arabica* and *C. canephora* cultivar/accessions

SNP ID	Flanking sequences and SNPs
Ca001	CCGCGAATCCTGTAAACCCAAAGCCACCCAATTCTCTCATGCACACTCACATTCCTTAC[A/C]ACTCTTTACCAGAATCAACAAAATCTCCAGGCTGCCATATTGTTTATGTTTATCGAGACCC
Ca007	TATTGAAGTGATTAATCGTGGAAGAGGAATGTCTGTTTTTGTTTTCTTGCAAATAAGCT[T/C]AATGTTGTGAGGTTTGCTGCAGATGCAGTTGGTTACGCAGTTGTGTTTGCCTGTTCATAGC
Ca008	CCCGTCTTCCACAAAAAGGCCAAAACTACAAAGAAGATTGTGCTAAGGCTGCAATGCCA[A/G]GGTTGTAAGCACGTTTCACAGCACCCTATCAAGAGGTGCAAGCACTTTGAAATTGGTGGTG
Ca009	TGGGTTTCTTGCGTTCGGTTTAGCCCGAATAATCTGCAGCCGACAATTGTTTCTGGGTC[T/G]TGGGATCGAACTGTGAAGATTTGGAATTTGAGTAACTGTAAGATGAGGGCAACTTTGGCTG
Ca012	CTAAAAATGCGAGGTGCTGGATCTGGAGCTGGAAGCTTACTTAAAGTTTTGGCCAACAA[C/A]TTTGATGTTCTTGCTTGGCCTGTGGTTAGTTTGGTGTATCCTCTTTATGCTTCCATCAGGG
Ca017	GAGTTGATGGAGCAGTTTCAGTTTTTGTCAGAAGAAGCTCTTCAACAGGAGTTCAACGA[A/T]CATGACTTTGTCTTGTAATTTACCACTTCTGTTCAACACAATTCTTGATCATTAATTGTAC
Ca022	GTGCCCATGCCTTGTATGTTCTTTGGTGGTGGGACTACTCAGTTCTTAACTAGTCGTGA[T/C]GGTGGAGGATGGATAGATGCTGCAAAGTTCTTAACAGGGGCATCAGCAGTGGGGAGCCTAG
Ca029	TTCCTGGTGAGTCCTGAAGATGGAAAAGATGGGGTTGGAAGGAGCGGAGTGTTGGAGGA[G/A]GTAATGAAGGGACTGTACTACGGAACTAAGGAGACCGTGGGTTGTGCTGCTGAGATGGTGA
Ca030	GGAGTGTTGGAGGAAGTAATGAAGGGACTGTACTACGGAACTAAGGAGACCGTGGGTTG[C/T]GCTGCTGAGATGGTGAAGAGGAATGCTGTTGAGATCGGGGACTTCAGATTCTTTGATGGAT
Ca032	GATGCTCAGCGAAGGCTGAGTAAATCCCATATACTTGTCAGTGGACTTACAGGCACTGT[C/T]GTTGAGTTCTGCAAGAACATTGTCCTTGCTGGAGTTGGTAGTTTGACATTGAATGATGATC
Ca034	GGATCTTTCAGCTGGGATCAGGACAATGAGAAAGTGAAGATTTATGCCTCTTTAGAGGG[G/A]GTTGACCAGGAGAAAATTGAGGCTGACTTTAAGCCGATGTCATTTGACCTCAAATTCCATG
Ca035	GCTTTCTCTCTCTCCAGATGCTTTTCTACTGTTCTGGAAGGCTTTAAGTATGCAAACTC[A/G]CATGAATGGGTAAAGCATGAAGGCCCGGTGGCTGCAGTTGGTATCACTGACCATGCTCAGG
Ca036	TGGTACGGCAAGTTTACGCCCACCCAGCGCTCCATTATTGTCGATTTTCTTCAATCCCT[T/A]AACTCCCCCAGGGCGGCTTCTCCCTCCGCCGCCTCCTGGTGGATGACGACCGAGAAGTACA
Ca040	CTCGTGCTATCTATTGCTGATGTTTATTGCCCTCCTAGAAAACGCTCACGTGTTAGCGC[A/G]CCATATGCTGTTGACAGTCGTTTGTTTAACAAAGAGCGGAACCCTTCCATTGAAACTCTTC
Ca048	TTTGAAATCAAATGTATGAATGACCCAAAAGCTTGCCTTCCTGGTTCCATTATTGTCAC[G/A]GCTACCAACTTTTGCCCTCCTAACAATGCACTCCCAAACAATGATGGAGGCTGGTGCAATC
Ca050	TCTCTATATGGCCTACCAAATTCCAAAACAAGACTAATGGGATTACTCCTCGCCGGTGG[A/C]TTCGGTTTTGTAGTCCTGAGCTTAGTCAAATAATAACCAAATGGTTAAAAACTGATAAATG
Ca055	CACACAACGTGTCCACAAAGAAGAGAAAAGAGATAGTGGAGCGTGCAGCTCAGCTAGAT[G/A]TTGTTGTTACTAACAAGCTTGCTAGGCTGCGGAGCCAGGAGGATGAATGAGCTTTTTGGCT
Ca056	GCGGAGCCAGGAGGATGAATGAGCTTTTTGGCTTTCATGCGCTGAACTTGGTTCATTTT[A/G]ATTATCATGCCTTTTTGTTTACGCTATCCCTTGTTAGTAACATTGTAATTTTGACAGATGA
Ca073	CGTAATATGTCTGTTATTGCACATGTTGATCATGGGAAGTCCACTCTTACTGATTCTCT[T/C]GTGGCTGCTGCTGGTATCATTGCTCAAGAAGTTGCTGGAGATGTTCGAATGACGGATACAA
Ca084	GGTTGTGGAAGTTGGACCTGAAGTAAAGAATTTGAAAGCTGGTGACAAAGTTGTGGCAT[A/T]TCTTAATCCTTTGTATGGTGGTGGATTGGCTGAGTTTGCTGTTGCCAAGGAGAGCTTGACT
Ca087	AATATCGCCACGCCTCATCACCATGAGGTTGGCTATCAGGGCTATGGGCAGCAGCACAG[C/A]ATTAATGGTGATGGGTATGGGAATCACCACAAGTACAATGACTACAACAGCCATGGCTATG
Ca092	TCCACCCTCTACGCCGTAGGCAGCCGTTCGGTGGAAAAAGCCTCAAACTTTGCGAAAGA[T/G]AATGGCTTTCCGGCTTCAGCAAAGGTATACGGCAGTTATGACGCCCTTCTAGATGACCCAG
Ca093	ATGGTTTTCATGGGTCCGACATTTTATCAACGCCTTACTCATATGGCTGAAGATAAAGT[C/A]AAATTTCGGAACACGGGACCAGTCCATCCACTCACTCGTCAGCCAGTGGCAGACAGGAAGC
Ca101	ACCACGTGCTGTCTTGTCCAGTCCAGATAATGATCAAATGATTGGAAGCAAAAACAAG[A/C]CAAAAGGTGATATACTTGCCAGTATGAAAAGGCAAAGTCTGTTTGAGAATAGACACGCCCGG
Ca102	GTTTGAGAATAGACACGCCCGGTGTAAGGTTACTCCGAGGCCTGTTGCTGCTGATGGCT[C/G]TATAAGCACAAGAACATCGCTTAAGGAAGTACCTGACGGTAAAGGTGATCTTCGAACTAGA
Ca103	GTGGAAGGCTACAGCCAGCATGACAATGAAGTCCAATATAGTCTCCTATTAGTTACTTA[A/T]AGGAATAAAGAGACTACTCATTTGAACTTCACAAATATGAACTTTATGATGTATTTTCTGT
Ca105	GTGTGCGTCTACCCATTGCGGAGCCGCGTATGCGTATGAAAAGAACCTGATCGACTCAC[A/T]TAATTCAGGCTCAGGATAGAGTTCCAGGAAGAAAAATGGTGAAATTGACTATGATTGCTCG
Ca106	ATGGGTCATGATGAATAAAGAGACTAGGAGACTGTCTAAAATTCCAGATGAGGTTCGAG[G/C]AGAAATAGAAGGCTATTATTTAGATTCACCTCCTATTGTGGATGAGGATGGCAGAAAGTTA
Ca107	CCAACAAAAGCAAGAAAGCATGAACGCCGAGGTTATTTCCAGCGGTCACCTGCAACTTT[A/T]CCCAACAAGTTCAAATACAAGACATTGTGGCTCCCAGCTTCCATTTAGCGGCAGGGCCAGT
Ca108	AGAGCAAAGACACCCTTCTCAAGCCAAGAGCCCCTGCTTCCATTCCTCTCTTGTCCCTC[A/C]ACGATGGACAAACTGAAGCTGTTTTCTACCGGTGCGGCTTTGGTTACAATTGTAACTATGT
Ca111	CATCTTCGTCAACTTCACATGACCAGTCACGACCCCGTAATGCAGGGTCAACTGGAAGA[G/A]CATCTGGTGCATCTACAACACAAACTCCTAGTGCAACATCTCTGCGATGGGATCGGCAAAC
Ca145	ACCGCGTTCGCCTGACCCACGCCCGCAAGAAGGGCGTTTACGAAGCACGCATGACCCCT[G/A]GCTGGGCGCTGTTTCGTGGCGAAGACCTGGTGCCCACCGCGCGTCTGGATCAGCAGGACGG

Table 3 (continued)

SNP ID	Flanking sequences and SNPs
Ca150	GCCCCAAAGATGCCTTAGTCGGCGGTTGGAGTAAGGCTGACCCCAAGGACCCAGAGGTG[C/G]TAGAGAACGGAAAATTTGCCATAGATGAGCACAACAAGGAGGCCGGTACCAAGTTGGAGTT
Ca155	CACTCAAAGGTGGGTGCTGATGAAGAAGAGGAGCCTGAGATAATCGAATCTGATGTTGA[C/G]CTTGATGACACTGAAGTTGTGGAGCCTGATAATGATCCTCCGCAACAGATGGGAGACCCTT
Ca169	TGGGACATGATGCTTCCAAACGAAGACCCTTTTAGAATCTTGGAGCACAGCCCTTTAAC[T/G]GTCCCCAAAGGGGTGGAGACGCTGGCCTTGGCACGCGCTGACTGGAAGGAGACGGCGAAGG
Ca171	CATTTGACATCACTTGATGGAGCTAAGGAAAGGCTTCAGTTGTACTCGGCAAACTTACT[G/A]GAAGAGGGATCGTTTGATGCAATAGTCGAGGGATGTGAAGGGGTTTTCCATACTGCATCTC
Ca175	TTTAACGATGTAATCGAGAAAATCTGTTGTGTCATCAAATTTGAACCCTCTGCTGATGG[G/A]GGTTCAATCTGCAAAACCACTAATACATACTACCCCAAAGGTGGTGCTCAGATCAGTGAGG
Ca176	GCCAGGGCCAGCCCTGCTCAAGCTAGCATGGTTGCACCCTTCACCGGCCTCAAAGCTGC[A/T]TCTTCTTTCCCCATTTCCAAGAAGTCCGTCGACATTACTTCCCTTGCCACCAACGGTGGAA
Ca178	TCAACCGCGTGAATGGCGGCCTGCAGTGGAAAATTGTTATTGGCACTCTCTATATCCTT[A/G]TCCTTGCAACTCAGGATTCTAAGGGCACATATACCGATTATGCAGTGGTTTTTGAGACCTT
Ca184	GCCTGAGGCAGTCCTTCAGACTGTTTCAAAGACCGGGAAGAAGACTTCTTTCTGGGAAG[G/A]AGGAGCATCAGCTGCACCTGAATCGAAGCCCGCAGAAACTGTTGCAGCTGCATAATTTGGG
Ca191	AAGTGCTTTCATTTTTGTGTCACCTCATGACTATCGTTTGGAATGGTGTTTTACACCTT[A/T]TGTGCGGAAAGTTGCATATCTTTGGTTACTCAGATAACGGATGAGGATGTTCAACAGCTAG
Ca194	AGTCGAAACTGATGCGGCGGTGACCAAATTCAAGAAAGTCATCTCTCTTCTAGGCCGAA[G/C]CAGAACTGGCCATGCTCGTTTTAGAAGAGGCCCCGTCCCCGTGGCTACAAATCCGGTTCCT
Ca197	GGACTTCACAATGGCAGGAATTTGATTGGAGGCATCAATAGCAAAAGGGCTTCAACATG[T/G]AAAGCAAATGCCTTCCCAGATTGGCCATTGATGGCAGTACTGGTTGAGCATGCTGAAGGAC
Ca206	GAGTTTGCTACTCGTCTGGGTAATGTCTTCATCATTGGAAAAGGTGCAAAACCCTGGGT[T/G]TCTCTTCCAAAGGGCAAAGGTATCAAGTTGTCAGTTATAGAGGAACAAAGGAAGAGGATTG
Ca214	TCTCTGTTTAACTTCATAAGCTGTCAAATATTTAACGTCACCGTCACCGTCAGTTCTGC[A/T]GAAAATGTCGCTGATTCCAAGCGTCTTCGGTGGCCGAAGAAGCAACGTTTTCGACCCATTT
Ca216	TCAGGGTGGTTAGTGCTGCAGTCTTTGATCTCAACGCAACGGCTCCTCCCTTCTTGTCG[G/A]CCACCATGCAGTTCACACTCACCACCAGGAACTCTAACCGGCGAGTCTCCTTTTTCTACGA
Ca222	ATATATTGCTTCTACTCTCTTCCCTTTTCCTTTCTCCCTTCTCCCCTTTTCCAAATTAA[C/G]TCCCGCTGATCATTTCTCTTTTCCAAATTCTCTCCTTCCTTTCTTAAATCCACCGCCCCCT
Ca225	AGGCACAAGATAGGAGAAGAGATTGGAGCGGTAGCCGCACTTGGAGCTGGTGGATTTGC[A/T]TTCCATGAGCATCACGAGAAGAAGGAAGCTAAAGAAGAAGAGGAAGAGGCCGAGGGGAAAA
Ca312	GCCAGAAATCCTGCTCGATGGCGAACATGGCATTGACAGGACCTTTGAGGTGGCTCTAA[C/T]AGTTTGGGCCGAGGTCTTCTTTTACCTTGCTGAGAACAATGTTCTGTTCGAGGGTATACTT
Ca346	ATTTCCAAGGCCATTGCTGAACATAAGCTCAGAATCATTTGCAACTGTAATTAACAGTG[A/G]AATCGGGAGAACCCTGGAGCCACCTTTTGATCCTTATGCTAATGACATCAACTATCTCATT
Ca356	TCCTATTGCCAACAAGTGCACAGACGAAACCTCAAGGAGGTGAGGTTGTAGCTGTTGGA[T/C]AGGGTCGCACAATTGGCAAGAGCAAGGTGGACATCAGTGTCAAGACTGGGACCCAAATTGT
Ca375	TCTTGCGTTTCTCCCAGTGGAAGCAGCAGGATTACACCGCGCGCCTGGTGCAAACCCCG[C/G]AAGCGCTGGCTGAGTTCCTCAAGCCGCTGTCTGACGCCGGCGTGGATATTTTCCACTGTTC
Ca396	CGCAACAGTTGAAAAATCCCAAGTTGAGGATGAAGGTGTCAATTTCTTATGATTTAGAT[T/C]ACCCTGATACTGAGAAGGAAGGGAAGAGTGATAAACAGGTTAAGAAGACCAAGAGGAAGCA
Ca397	AACTTTAAAATCGTGGCTACTGAGATTGACGAGGATAAGCAGACCGAGAAGGACAGATG[G/A]AAGGGCCTAAGCACTGATACCTCTGATGATCAACAAGACATCACCAGAGGAAAGGGCATGG
Ca398	GGGACTCGTTAAGAGGGAGGAGCTTTTCATTACTACCAAGCTGTGGAATTCAGACCATG[T/C]CCACGTTCTCGAGGCTTGCAAAGACAGCCTGAAAAAGCTTCGTCTTGATTATCTTGACCTG

compared with the SNPs recently discovered by Moncada et al. (2016), based on a F2 population derived from Catuai x CCC1046, no overlapping was found between these two sets of data. Since most of the polymorphism in Moncada's et al. (2016) mapping population was likely contributed by the wild *C. arabica* (CCC1046), this part of genetic diversity might have not been captured in our pipeline of SNP identification based on a relatively small number of genotypes. The biased sampling of species, populations or individual genotypes in SNP discovery will have important impact when the developed SNP markers are used in studies of population genetics, molecular systematics and genetic diversity analysis, because a small sample size is more likely to identify common alleles rather than rare alleles. Nonetheless, this potential ascertainment bias in SNPs is less a concern when SNP markers are used by curators and breeders for routine management of coffee germplasm, including genotype identification, parentage/sibship analysis, and assignment of individual to a known population.

Coffee Cultivar Identification

Despite the low polymorphic rate detected in *C. arabica*, the SNP panel used was still able to reliably identify the tested *C.*

Table 4 Minor allele frequency, information index, observed and expected heterozygosity, and inbreeding coefficient of the 55 SNP loci scored on 25 coffee accessions

SNP ID	Minor allele frequency	Information index	Observed heterozygosity	Expected heterozygosity	Inbreeding coefficient
Ca001	0.140	0.405	0.280	0.241	−0.163
Ca007	0.460	0.690	0.600	0.497	−0.208
Ca008	0.440	0.653	0.560	0.461	−0.215
Ca009	0.360	0.686	0.880	0.493	−0.786
Ca012	0.140	0.405	0.280	0.241	−0.163
Ca017	0.280	0.593	0.560	0.403	−0.389
Ca022	0.200	0.500	0.400	0.320	−0.250
Ca029	0.440	0.686	0.160	0.493	0.675
Ca030	0.440	0.686	0.160	0.493	0.675
Ca032	0.200	0.500	0.400	0.320	−0.250
Ca034	0.326	0.631	0.478	0.440	−0.088
Ca035	0.460	0.690	0.600	0.497	−0.208
Ca036	0.440	0.686	0.880	0.493	−0.786
Ca040	0.080	0.279	0.160	0.147	−0.087
Ca048	0.400	0.673	0.640	0.480	−0.333
Ca050	0.080	0.279	0.160	0.147	−0.087
Ca055	0.140	0.405	0.120	0.241	0.502
Ca056	0.440	0.686	0.880	0.493	−0.786
Ca073	0.440	0.686	0.080	0.493	0.838
Ca084	0.240	0.551	0.400	0.365	−0.096
Ca087	0.360	0.653	0.160	0.461	0.653
Ca092	0.280	0.593	0.560	0.403	−0.389
Ca093	0.280	0.593	0.560	0.403	−0.389
Ca101	0.260	0.573	0.120	0.385	0.688
Ca102	0.220	0.527	0.440	0.343	−0.282
Ca105	0.271	0.584	0.542	0.395	−0.371
Ca106	0.120	0.367	0.160	0.211	0.242
Ca107	0.080	0.279	0.080	0.147	0.457
Ca108	0.480	0.692	0.080	0.499	0.840
Ca111	0.460	0.690	0.600	0.497	−0.208
Ca150	0.120	0.367	0.160	0.211	0.242
Ca155	0.340	0.641	0.600	0.449	−0.337
Ca169	0.320	0.627	0.640	0.435	−0.471
Ca171	0.300	0.611	0.360	0.420	0.143
Ca175	0.040	0.168	0.080	0.077	−0.042
Ca176	0.460	0.690	0.040	0.497	0.919
Ca178	0.060	0.227	0.120	0.113	−0.064
Ca184	0.100	0.325	0.200	0.180	−0.111
Ca191	0.360	0.653	0.160	0.461	0.653
Ca194	0.180	0.471	0.360	0.295	−0.220
Ca197	0.040	0.168	0.080	0.077	−0.042
Ca206	0.280	0.593	0.240	0.403	0.405
Ca214	0.060	0.227	0.120	0.113	−0.064
Ca216	0.180	0.471	0.200	0.295	0.322
Ca222	0.300	0.611	0.200	0.420	0.524
Ca225	0.480	0.692	0.080	0.499	0.840
Ca103	0.340	0.641	0.600	0.449	−0.337

Table 4 (continued)

SNP ID	Minor allele frequency	Information index	Observed heterozygosity	Expected heterozygosity	Inbreeding coefficient
Ca145	0.440	0.686	0.880	0.493	−0.786
Ca312	0.360	0.653	0.480	0.461	−0.042
Ca346	0.200	0.500	0.160	0.320	0.500
Ca356	0.460	0.690	0.120	0.497	0.758
Ca375	0.380	0.664	0.600	0.471	−0.273
Ca396	0.040	0.168	0.080	0.077	−0.042
Ca397	0.460	0.690	0.920	0.497	−0.852
Ca398	0.040	0.168	0.080	0.077	−0.042
Mean	0.281	0.536	0.351	0.362	0.025

arabica cultivars (Figs. 1b and 2). Some cultivars only differed by a single SNP marker, but the differences are robust and fully repeatable, as examined in the present experiment by genotyping three different times. Results from multiple samples (independent DNA extractions) of the same cultivar showed 100 % concordance, demonstrating that the SNP panel is a reliable tool for generating coffee DNA fingerprints with high accuracy, in spite of the high concentration of polyphenonic compound and polysaccharide content in the coffee leaf samples. The SNP profiles enabled the differentiation of Puerto Rican cultivars Limaní, Frontón, and TARS 18087 from the rest of the *C. arabica* cultivars. The unique status of Limaní and Frontón is expected, because both Limaní and Frontón were Puerto Rican selections from hybrid progenies of *C. arabica* and Timor coffee. Specifically, Limaní was selected from the Sarchimor population (Sarchi x Timor) developed by the Coffee Rust Research Center in Portugal (Várzea et al. 2009), whereas Frontón was a selection from the Catimor population. The cultivar name of TARS 18087 was unknown and the present result suggested that this accession was a hybrid of *C. arabica* and Timor coffee too, based on its proximity with Limaní.

The cultivar Borbón and synonymous group #1 (including Caturra, and Catuai) only differed by a single SNP marker (Ca171), indicating the origin of Caturra and Catuai as mutants or offspring derived from Borbón. The 121 bp sequence of Ca171 showed 100 % homology with 10 ESTs in NCBI, most of which were expressed in root and calli cDNA libraries (Vieira et al. 2006; Mondego et al. 2011). A search of the corresponding SNPs and their flanking sequences in GenBank showed that the putative gene codes for cinnamoyl-CoA reductase, which is one of the enzymes converting phenolic acids into the monomeric units of lignins, a main class of structural materials in the support tissues of vascular plants (Gross 1980). Lignification is important in the vascular plant cell wall, because it enables xylems to withstand the negative pressure generated during water transport. This sequence also corresponds to bifunctional dihydroflavonol 4-reductase/flavanone 4-reductase (DFR), as annotated in the Coffee Genome Sequencing Project (Supplemental Table 1). DFR was reported as a key element of regulation for flavonoid and anthocyanin biosynthesis in olive (Martinelli and Tonutti 2012) and peach (Tsuda et al. 2004). Another example of a polymorphic SNP is Cac35, The contig that contained Ca35 was deduced to produce the H-protein in the glycine cleavage system H protein, and the homologous gene annotated in the Coffee Genome Sequencing Project (Supplemental Table 1). The glycine cleavage system H-protein is one of the four different component proteins required for photorespiration. It was reported that the H-protein occurred specifically in leaf tissues such as Pea (Macherel et al 1990). The H-protein was also reported to have differential responses on tomato plants treated with jasmonic Acid and salicylic Acid (Afroz et al. 2010). Therefore, the SNP variation (including mutation) among the closely related *C. arabica* cultivars may have reflected a functional mutation in terms of adaptability to biotic and/or abiotic stresses, or may have been selected for different qualities attributes.

The present study also revealed several discrepancies regarding the documented pedigree relationship of Arabica cultivars. For example, cultivar Catuai is a hybrid progeny between Caturra and Mundo Nuevo. However, the 55 SNP markers could not differentiate Caturra and Catuai, indicating these two cultivars belong to the same synonymous group. In contrast, Pacas is considered a mutant of Borbón, but the two cultivars showed differences in three SNP loci, which may not be easily explained by mutation. These discrepancies might be plausibly explained by the occurrence of mislabeling in the sampled Puerto Rican coffee cultivars, especially since the present study used only one accession from each of the *C. arabica* cultivars, which may have biased the sample representation. It's also possible that the hypothesized pedigree relationship among *C. arabica* cultivars cannot be fully supported by molecular evidence. Systematic genotype identification in *C. arabica* coffee germplasm is still needed. Additional samples for each cultivar from Puerto Rico, as well as from other *Coffea* genebanks,

Table 5 DNA fingerprints based on the array of 55 SNPs for coffee genotype identification, showing truncated profiles of 17 loci

Accessions	Ca001	Ca007	Ca008	Ca012	Ca155	Ca032	Ca034	Ca035	Ca040	Ca050	Ca056	Ca073	Ca102	Ca171	Ca194	Ca216	Ca346
Limaní	A A	C T	A G	A C	C G	C T	A G	A A	A A	A C	A G	C T	C G	A G	C G	A A	A A
Frontón	A A	T T	G G	C C	G G	C T	A G	A G	A A	A C	A G	C T	C G	A G	C G	A G	A A
Puerto Rico	A A	C T	A G	A C	C G	C T	A G	A G	A A	C C	A G	T T	C G	A G	C C	A A	A G
Catuai (synonymous group 1)	A A	C T	A G	A C	C G	C T	A G	A G	A A	C C	A G	T T	C G	A G	C C	A G	A A
Caturra (synonymous group 1)	A A	C T	A G	A C	C G	C T	A G	A G	A A	C C	A G	T T	C G	A G	C C	A G	A A
TARS 18087	A A	C T	A G	A C	C G	T T	A G	A G	A A	A C	G G	T T	C G	A G	C G	A G	A A
Borbón (synonymous group 2)	A A	C T	A G	A C	C G	C T	A G	A G	A A	C C	A G	T T	C G	A A	C C	A G	A G
Sumatra (synonymous group2)	A A	C T	A G	A C	C G	C T	A G	A G	A A	C C	A G	T T	C G	A A	C C	A G	A G
Mundo Nuevo	A A	C T	A G	C C	C G	C T	A A	A G	A A	C C	A G	T T	C G	A G	C C	A G	A G
Enrea	A A	C T	A G	C C	C G	C T	A A	A G	A A	C C	A G	T T	C G	A G	C C	A G	A A
Harrar Rojo	A A	C T	A G	C C	C G	T T	A G	A A	A A	C C	A G	T T	C C	A G	C C	A G	A A
Pacas	A A	C T	G G	C C	C G	C T	A G	A G	A A	C C	G G	T T	C C	A A	C C	A G	A A
Robusta EEA	A A	T T	A A	C C	C G	T T	G G	A G	A A	C C	G G	C C	C C	G G	C G	A G	A A
A5 00726	A A	C C	A A	C C	C G	T T	G G	A A	A G	C C	A G	C C	C C	G G	C G	A G	A A
A5 00738	A C	C T	A A	C C	G G	T T	G G	G G	A A	C C	A G	C C	C C	G G	C C	G G	A G
A5 00742	A C	C C	A A	C C	C G	T T	A G	A A	A A	C C	A G	C C	C C	G G	C G	G G	A A
A5 00797	A C	T T	A A	C C	G G	T T	G G	G G	A A	C C	A G	C C	C C	G G	C C	G G	A G
A5 00906	A C	C C	A A	C C	C G	T T	G G	G G	A A	C C	A G	C C	C C	G G	C C	G G	A A
B4 00139	A C	C T	A G	C C	G G	T T	G G	A G	A A	C C	A G	C C	C C	G G	C G	G G	A G
B4 00140	A A	T T	A A	C C	G G	T T	G G	A G	A A	C C	A G	C C	C C	G G	C C	A G	A A
B4 00262	A C	C C	A A	C C	G G	T T	G G	A G	A G	C C	A G	C C	C C	G G	C C	G G	A A
B4 00265	A A	T T	A G	C C	G G	T T	G G	A A	A A	C C	A G	C C	C C	G G	C G	G G	A G
B4 00268	A A	C T	A G	C C	C C	T T	G G	A G	A A	A C	A G	C C	C C	G G	C G	G G	A A
Robusta #69	A A	T T	A A	C C	G G	T T	G G	A G	A G	C C	G G	C C	C C	G G	C C	G G	A A
Robusta #70	A C	C T	A G	C C	G G	T T	G G	A A	A G	C C	A G	C C	C C	G G	C C	G G	G G

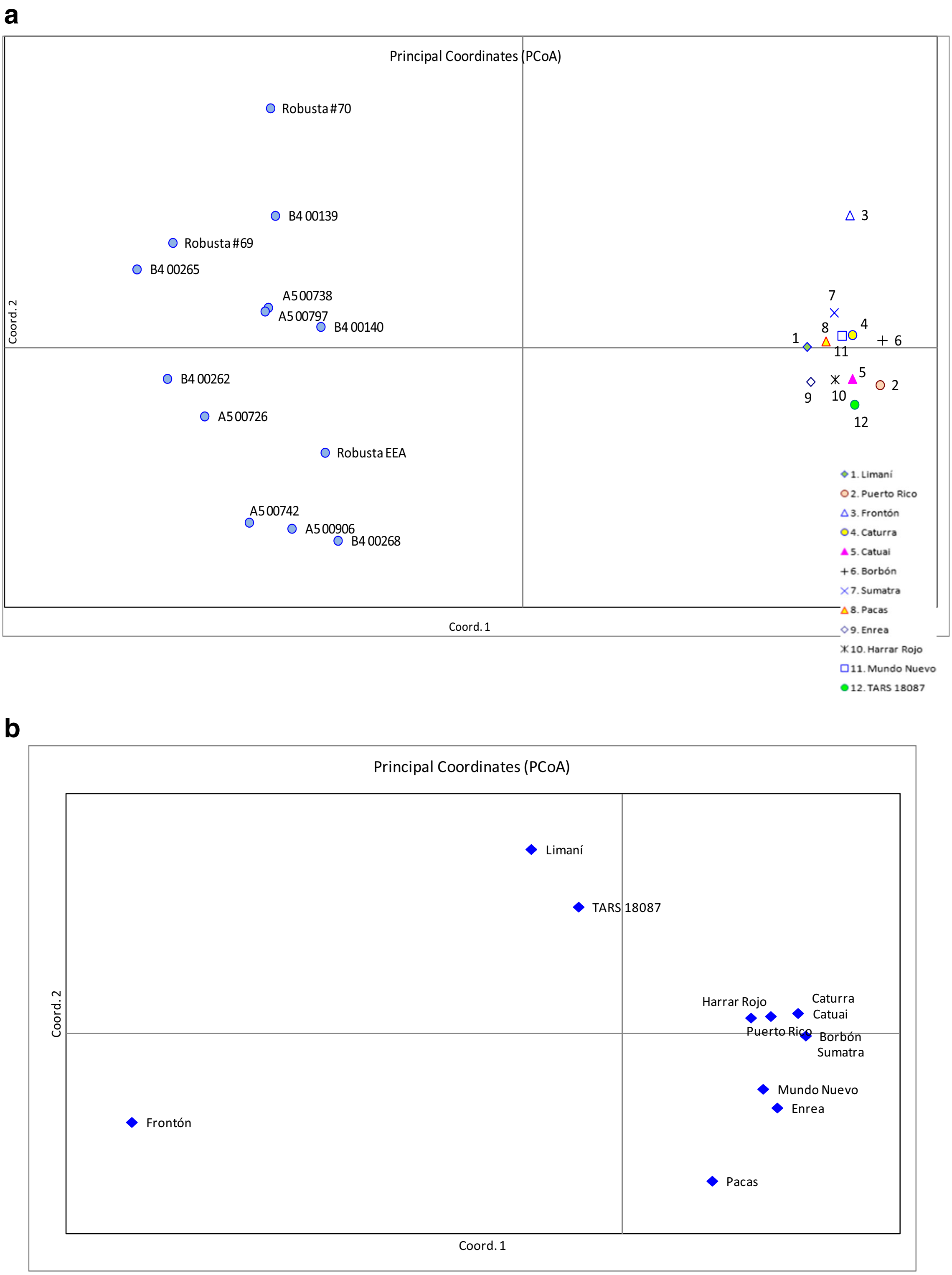

Fig. 1 **a** PCoA plot of 25 coffee accessions including 12 *C. arabica* and 13 *C. canephora* (Robusta) accessions from Puerto Rico. The plane of the first three main PCOA axes accounted for 71.98 % of the total variation (first axis = 57.49 %, second = 8.35 % and third = 6.14 %). **b** PCoA plot of 12 *C. arabica* accessions from Puerto Rico. The plane of the first three main PCOA axes accounted for 65.5 % of the total variation (first axis = 28.9 %, second = 22.1 % and third = 14.5 %)

need to be examined to clarify the validity of the hypothesized pedigree relationships among the Arabica cultivars. The advantage of SNP markers in genotyping accuracy will enable the identification of mutation groups, which, in contrast, cannot be easily detected by markers such as SSRs, due to their larger error rate in genotyping and allele calling.

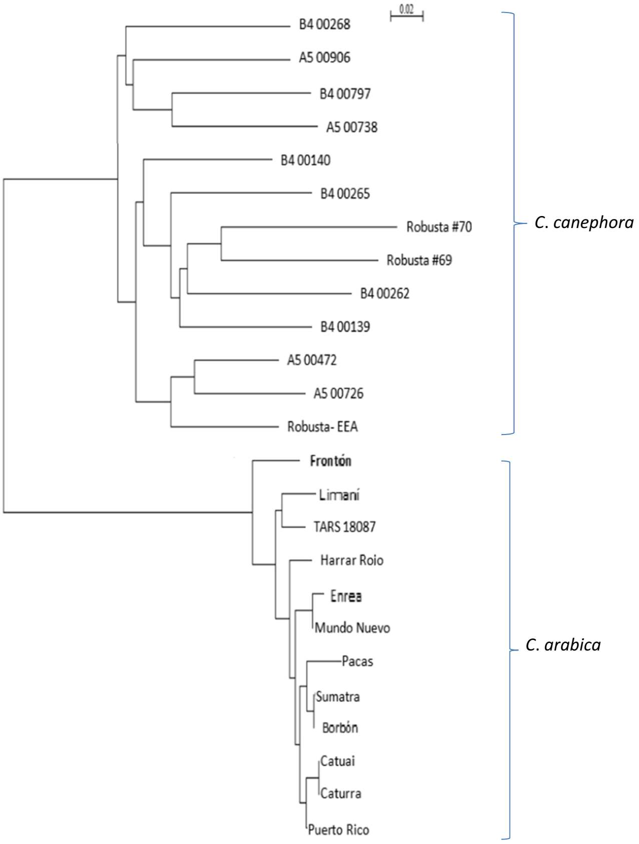

Fig. 2 Neighbor-joining unrooted tree depicting the relationship among the 25 coffee accessions from Puerto Rico. Identification of accessions corresponds to samples listed in Table 1

A much larger SNP variation was found in *C. canephora*, in which 42 SNP markers (out of the total of 55 for both species) were found to be polymorphic (Fig. 1a and b). Individual genotype matching (pairwise comparisons) based on the 48 SNP markers showed that each of the 13 accessions has a unique SNP profile. The average distance among the 13 *C. canephora* accessions is 31.58, whereas the average genetic distance was 6.86 among the 12 *C. arabica* cultivars. This level of polymorphism suggested that this set of SNPs can provide sufficient differentiation power for the purpose of *C. canephora* genotype identification. A *C. canephora* tree given by the French government was introduced to Puerto Rico in 1906, and it "fruited heavily in some seasons" (McClelland 1924). It is not known whether progeny from this tree were subsequently planted in the field. In 1914 and 1915 *C. canephora* seeds were obtained from the Java Experiment Station at Buitenzorg, Indonesia and planted in Puerto Rico (McClelland 1924). However, based on McClelland's report (1924), it is not clear if the Java introductions were a single accession or a mixture of different accessions. Reference *C. canephora* accessions from Java need to be compared in the follow-up studies in order to clarify the genetic identity of the *C. canephora* germplasm in Puerto Rico for the purposes of germplasm conservation.

In conclusion, we conducted a pilot study on the development of SNP markers for coffee and employed them for varietal genotyping, using a nanofluidic array. This technology enabled us to generate high quality SNP profiles, which can serve as universal DNA fingerprints for cross-laboratory genotype comparisons. The SNP panel provides a useful tool to assist in coffee germplasm management, quality control of planting material propagation, and protection of varietal rights in the international coffee community. Additional efforts to develop and validate more SNP markers are underway in order to develop a high quality genotyping panel of SNPs for cultivar identification and genetic diversity analysis in coffee. This information will have a significant potential for practical application.

Acknowledgments We would like to give special thanks to Stephen Pinney (USDA, ARS) for assistance with SNP genotyping using the nanofluidic array system. References by USDA to a company and/or product are only for the purposes of information and do not imply approval or recommendation of the product to the exclusion of others that may also be suitable.

Compliance with Ethical Standards

Conflict of Interest The authors declare no conflict of interest.

References

Afroz A, Khan MR, Komatsu S (2010) Determination of proteins induced in response to jasmonic acid and salicylic acid in resistant and susceptible cultivars of tomato. Protein Pept Lett 17:1–11

Anthony F, Astorga C, Avendaño J et al (2007) Conservation of coffee genetic resources in the CATIE field genebank. In: Engelmann F, Dulloo ME, Astorga C et al (eds) Conserving coffee genetic resources. Bioversity International, Rome, pp 23–34

Anthony F, Bertrand B, Quiros O et al (2001) Genetic diversity of wild coffee (*Coffea arabica* L.) using molecular markers. Euphytica 118:53–65

Combes MC, Dereeper A, Severac D et al (2013) Contribution of subgenomes to the transcriptome and their intertwined regulation in the allopolyploid *Coffea arabica* grown at contrasted temperatures. New Phytol 200:251–260

Cubry P, Musoli P, Legnaté H et al (2008) Diversity in coffee assessed with SSR markers: structure of the genus *Coffea* and perspectives for breeding. Genome 2008(51):50–63

Davis AP (2010) Six species of *Psilanthus* transferred to *Coffea* (Coffeeae, Rubiaceae). Phytotaxa 10:41–45

Davis AP (2011) *Psilanthus mannii*, the type species of *Psilanthus*, transferred to *Coffea*. Nordic J Bot 29:471–472

Davis AP, Govaerts R, Bridson DM et al (2006) An annotated taxonomic conspectus of the genus *Coffea* (Rubiaceae). Bot J Linn Soc 152:465–512

de Kochko A, Akaffou S, Andrade AC et al (2010) Advances in *Coffea* genomics. Adv Bot Res 53:23–63

Denoeud F, Carretero-Paulet L, Dereeper A et al (2014) The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 345:1181–1184

Dereeper A, Bocs S, Rouard M et al (2015) The coffee genome hub: a resource for coffee genomes. Nucleic Acids Res 43:D1028–D1035

Dieringer D, Schlötterer C (2003) Microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. Mol Ecol Notes 3:167–169

Fang W, Meinhardt LW, Tan H et al (2014) Varietal identification of tea (*Camellia sinensis*) using nanofluidic array of single nucleotide polymorphism (SNP) markers. Hortic Res 1:14035

FAOSTAT (2014) Food and Agriculture Organization of the United Nations, Statistics Division. Available online at: http://faostat.fao.org/

Felsenstein J (1989) Mathematics vs. evolution: mathematical evolutionary theory. Science 246:941–942

Fluidigm (2011) Fluidigm SNP Genotyping User Guide Rev H1, PN 68000098. South San Francisco, CA: Fluidigm Corporation. http://www.mscience.com.au/upload/pages/fluidigmtech/fluidigm-snp-genotyping-user-guide-151112.pdf

Geleta M, Herrera I, Monzón A et al (2012) Genetic diversity of Arabica coffee (*Coffea arabica* L.) in Nicaragua as estimated by simple sequence repeat markers. Sci World J 2012:939820

Gole TW, Denich M, Teketay D, Vlek PLG (2002) Human impacts on the *Coffea arabica* genepool in Ethiopia and the need for its *in situ* conservation. In: Engels JMM, Ramanatha Rao V, Brown AHD, Jackson MT (eds) Managing plant genetic diversity. CABI Publishing, Oxon, pp 237–247

Gross GG (1980) The biochemistry of lignification. Adv Bot Res 8:26–63

Hamon P, Hamon S, Razafinarivo NJ et al (2015) *Coffea* genome organization and evolution. In: Preedy VR (ed) Coffee in health and disease prevention. Academic, San Diego, pp 29–37

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877

International Coffee Organization (2014) World coffee trade (1963–2013): a review of the markets, challenges and opportunities facing the sector. International Coffee Organization, London, ICC 111–5 Rev.1.29

Ji K, Zhang D, Motilal L et al (2013) Genetic diversity and parentage in farmer varieties of cacao (*Theobroma cacao* L.) from Honduras and Nicaragua as revealed by single nucleotide polymorphism (SNP) markers. Genet Resour Crop Evol 60:441–453

Krishnan S, Ranker TA (2012) Coffee genomics. In: Benkeblia N (ed) OMICs technologies: tools for food science. CRC Press, Boca Raton, pp 227–247

Lashermes P, Paczek V, Trouslot P et al (2000) Single-locus inheritance in the allotetraploid *Coffea arabica* L. and interspecific hybrid *C. arabica x C. canephora*. J Hered 91:81–85

Leroy T, De Bellis F, Legnate H, Musoli P, Kalonji A, Loor Solorzano RG, Cubry P (2014) Developing core collections to optimize the management and the exploitation of diversity of the coffee *Coffea canephora*. Genetica 142:185–99

Lewin B, Giovannucci D, Varangis P (2004) Coffee markets: new paradigms in global supply and demand. The International Bank for Reconstruction and Development, Agriculture and Rural Development Discussion Paper 3, 150 pp

Liu W, Xiao ZD, Bao XL et al (2015) Identifying litchi (*Litchi chinensis* Sonn.) cultivars and their genetic relationships using single nucleotide polymorphism (SNP) markers. PLoS ONE 10(8), e0135390

López-Gartner G, Cortina H, McCouch SR et al (2009) Analysis of genetic structure in a sample of coffee (*Coffea arabica* L.) using fluorescent SSR markers. Tree Genet Genomes 5:435–446

Macherel D, Lebrun, Gagnon M, Neuburger M, Douce R (1990) cDNA cloning, primary structure and gene expression for H-protein, a component of the glycine-cleavage system (*glycine decarboxylase*) of pea (Pisum sativum) leaf mitochondria. Biochem J 268:783–789

Martinelli F, Tonutti P (2012) Flavonoid metabolism and gene expression in developing olive (Olea europaea L.) fruit. Plant Biosyst - Int J Dealing Asp Plant Biol 146(sup1):164–170. doi:10.1080/11263504.2012.681320

Masumbuko LI, Bryngelsson T (2006) Inter simple sequence repeat (ISSR) analysis of diploid coffee species and cultivated *Coffea arabica* L. from Tanzania. Genet Resour Crop Evol 53:357–366

McClelland TB (1924) Coffee varieties in Porto Rico. Porto Rico Agricultural Experiment Station, Mayaguez, Porto Rico, Bulletin No. 30, 27 pp

Missio RF, Caixeta ET, Zambolim EM et al (2010) Polymorphic information content of SSR markers for *Coffea* spp. Crop Breed Appl Biotechnol 10:89–94

Moncada P, McCouch S (2004) Simple sequence repeat diversity in diploid and tetraploid *Coffea* species. Genome 47:501–509

Moncada P, Tovar E, Montoya JC, González A, Spindel J, McCouch S (2016) A genetic linkage map of coffee (*Coffea arabica* L.) and QTL for yield, plant height, and bean size. Tree Genet Genomes 12:5

Mondego JMC, Vidal RO, Carazzolle MF et al (2011) An EST-based analysis identifies new genes and reveals distinctive gene expression features of *Coffea arabica* and *Coffea canephora*. BMC Plant Biol 11:30

Monroig Inglés MF (n. d.) Descripción de variedades de *Coffea arabica* más cultivadas en Puerto Rico. Available online: http://academic.uprm.edu/mmonroig/id45.htm

Moore PH, Ming R (eds) (2008) Genomics of tropical crops plants. Springer Science + Business Media, New York, **581 pp**

Nijveen H, van Kaauwen M, Esselink DG et al (2013) QualitySNPng: a user-friendly SNP detection and visualization tool. Nucleic Acids Res 41:W587–W590

Osorio N (2002) The global coffee crisis: a threat to sustainable development. International Coffee Organization, London. Available online: http://www.ico.org/documents/globalcrisise.pdf

Peakall ROD, Smouse PE (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research - an update. Bioinformatics 28:2537–2539

Perrière G, Gouy M (1996) WWW-query: an on-line retrieval system for biological sequence banks. Biochimie 78:364–369

Razafinarivo NJ, Guyot R, Davis AP et al (2013) Genetic structure and diversity of coffee (*Coffea*) across Africa and the Indian Ocean islands revealed using microsatellites. Ann Bot 111:229–248

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Tang JF, Vosman B, Voorrips RE et al (2006) QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. BMC Bioinform 7:438

Tshilenge P, Nkongolo KK, Mehes M et al (2009) Genetic variation in *Coffea canephora* L. (var. Robusta) accessions from the founder gene pool evaluated with ISSR and RAPD. Afr J Biotechnol 8:380–390

Tsuda T, Yamaguchi M, Honda C, Moriguchi T (2004) Expression of anthocyanin biosynthesis genes in the epicarp of peach and nectarine fruit. J Am Soc Hortic Sci 129:857–862

Várzea VMP, Marques VD, Pereira AP, Silva MC (2009) The use of Sarchimor derivatives in coffee breeding resistance to leaf rust. 22nd International Conference on Coffee Science, ASIC 2008, Campinas, SP, Brazil, 14–19 September, 2008 pp. 1424–1429

Vega FE (2008) The rise of coffee. Am Sci 96:138–145

Vega FE, Ebert A, Ming R (2008) Coffee germplasm resources, genomics, and breeding. Plant Breed Rev 30:415–447

Vidal RO, Mondego JMC, Pot D et al (2010) A high-throughput data mining of single nucleotide polymorphisms in *Coffea* species expressed sequence tags suggests differential homeologous gene expression in the allotetraploid *Coffea arabica*. Plant Physiol 154:1053–1066

Vieira LGE, Carvalho AA, Augusto CC et al (2006) Brazilian coffee genome project: an EST-based genomic resource. Braz J Plant Physiol 18:95–108

Vieira ESN, Von Pinho EVR, Carvalho MGG et al (2010) Development of microsatellite markers for identifying Brazilian *Coffea arabica* varieties. Genet Mol Biol 33:507–514

Wang B, Tan H, Fang W et al (2015) Developing single nucleotide polymorphism (SNP) markers from transcriptome sequences for identification of longan (*Dimocarpus longan*) germplasm. Hortic Res 2: 14065

Wu GA, Prochnik S, Jenkins J et al (2014) Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. Nat Biotechnol 32:656–662

Yuyama PM, Júnior OR, Ivamoto ST et al (2016) Transcriptome analysis in *Coffea eugenioides*, an Arabica coffee ancestor, reveals differentially expressed genes in leaves and fruits. Mol Genet Genomics 291:323–336

Zhang D, Mischke S, Goenaga R et al (2006) Accuracy and reliability of high-throughput microsatellite genotyping for cacao individual identification. Crop Sci 46:2084–2092