

# Genetic diversity and population structure of a Peruvian *Coffea arabica* L. collection

Roberto Mansilla-Samaniego<sup>1\*</sup>, Rosa Espejo-Joya<sup>1</sup>, Giovanni Bernacchia<sup>2</sup>, Javier Wither-Villavicencio<sup>1</sup>, Cinthia Quispe-Apaza<sup>1</sup>, and César López-Bonilla<sup>1</sup>

<sup>1</sup>Universidad Nacional Agraria La Molina, Av. La Molina s/n, La Molina, 15024 Lima, Perú.

\*Corresponding author (rmansilla@lamolina.edu.pe).

<sup>2</sup>Università degli Studi di Ferrara, Dip Scienze della Vita e Biotecnologie, Via Borsari 46, 44121 Ferrara, Italia.

Received: 23 September 2020; Accepted: 21 December 2020; doi:10.4067/S0718-58392021000200138

## ABSTRACT

Peru is an important producer of specialty coffee beans in the world, however the genetic of their coffee plant populations is unknown. Therefore, the genetic diversity and population structure of a Peruvian germplasm collection of plant coffee was analyzed to ascertain its potential use in plant breeding and conservation. In this work 54 DNA genotypes from 17 *Coffea arabica* L. and one *C. canephora* Pierre ex A. Froehner accessions were analyzed by microsatellite and sequence-related amplified polymorphism (SRAP) markers. In the assessment of molecular markers both systems were adequate to perform *C. arabica* germplasm collection genetic analysis. The obtained genetic diversity estimators were similar to germplasm assessments from other plant breeding programs. In the population structure analysis, the genetic differentiation ( $G_{ST} = 0.6584$ ) and genetic flow ( $N_m = 0.2594$ ) estimators were high. In analysis of molecular variance (AMOVA), total variation was divided 43.05% between accessions and 56.95% within accessions. In the Bayesian analysis with STRUCTURE software using admixed model for  $K = 2$ , the *C. arabica* and *C. canephora* accessions were separated, while for  $K = 7$ , the *C. arabica* accessions were grouped similarly to what obtained by the unweighted pair group method with arithmetic mean (UPGMA) dendrogram. The high genetic differentiation and genetic structuring of the accessions would indicate that the cultivars, from which the accessions were originated, have been preserved over time. The genetic diversity of Peruvian coffee might be the consequence of a long history of introductions of cultivars from different origins.

**Key words:** Coffee, genetic diversity, microsatellites, molecular markers, Peruvian coffee, SRAP.

## INTRODUCTION

*Coffea arabica* L. is a shrubby autogamous species originating from the southeast of Ethiopia and the north of Kenya on the African continent. It derived from the hybridization between *Coffea canephora* Pierre ex A. Froehner and *Coffea eugenioides* S. Moore, therefore its genome is the fusion of the two genomes (Romero et al., 2010). The *C. arabica* cultivars are responsible for 70% of the coffee trade worldwide; while the remaining 30% comes from *C. canephora* and others related species. In addition, the main cultivars grown in coffee-producing countries, and particularly in the Americas, have a narrow genetic pool, because their propagation started from very few plants that belonged to the Typica and Bourbon cultivars (Aerts et al., 2012). Accordingly, the Arabica cultivars that are distributed worldwide rose from the mutation and hybridization of those materials (Van Der Vossen et al., 2015). However, in the 1960s coffee genetic improvement programs performed interspecific crossbreeding among traditional cultivars of *C. arabica* and *C. canephora* to introduce new traits into the crop, such as pest and disease tolerance, especially leaf rust caused by *Hemileia vastatrix*.

For this reason, a spontaneous hybrid between *C. arabica* and *C. canephora*, discovered on the Timor Island in Southeast Asia, was used as an intermediary (Romero et al., 2014). These programs have increased the genetic diversity of *C. arabica* by crossbreeding traditional cultivars and Timor hybrid descendants therefore introducing genes of interest along with unwanted traits.

Peru is an important producer of specialty coffee beans in the world. However, when and how did coffee cultivars arrive in Peru? It is still unclear. According to the Junta Nacional de Café, the first plants arrived in the mid-18th century and were sown in the province of Chinchao (Huanuco Region). Shortly after, others cultivars arrived from Guayaquil, Ecuador, and were taken to the forest in central Peru. However, at different times throughout the history of coffee cultivation in Peru, improved genotypes might have arrived from neighboring coffee producing countries without being registered. Furthermore, according to Palomino et al. (2014) a mixture of cultivars is believed to exist. Despite the long history as coffee producer, Peru has no records of any genetic improvement program similar to those implemented in Colombia, Brazil and other countries in Central America. However, it is known that the coffee grown in the Peruvian forest produces grains with high organoleptic qualities, which are an important requirement for a progenitor in genetic improvement programs. Therefore, knowing the genetic potential of genotypes in a germplasm collection is relevant in order to perform efficient crossings.

Molecular markers have been very useful to perform genetic diversity and population structure analysis of germplasm collections in most cultivated species. In *C. arabica* crop the genetic analysis has been performed with molecular markers such as random amplified polymorphic DNA (RAPD), inter-simple sequence repeats (ISSR) (Yan et al., 2019), amplified fragment length polymorphism (AFLP), restriction fragment length polymorphisms (RFLP), simple sequence repeat (SSR or microsatellites) and expressed sequence tag-derived simple sequence repeats (EST-SSRs). Microsatellites, despite showing some limits in their use in tetraploid species such as coffee, have been considered very efficient to determine the genetic diversity, population structure and germplasm characterization in coffee collections (Vieira et al., 2010; Teresa et al., 2010; Geleta et al., 2012; Al-Murish et al., 2013; Mishra et al., 2014; Motta et al., 2014; Ferrao et al., 2015; Alemayehu, 2017; Vieira et al., 2017). Microsatellites have also been shown to be useful in linkage mapping (Pestana et al., 2015). Sequence-related amplified polymorphism (SRAP) markers have recently been used efficiently in studies of genetic diversity and population structure in germplasm collections of coffee (Al-Murish et al., 2013; Mishra et al., 2014; Jingade et al., 2019; Yan et al., 2019). These are characterized by a forward primer containing a GGCC sequence close to the 3' terminus, which preferably aligns with GC-rich sequences in coding regions. On the contrary, the reverse primer contains an AATT sequence, which preferably aligns with intron sequences. In addition, SRAP markers are cheap when compared to others molecular markers, require little amount of DNA, their band patterns are unambiguous and highly reproducible (Jingade et al., 2019). Consequently, in this work the genetic diversity and population structure of a Peruvian collection coffee germplasm were analyzed using microsatellite and SRAP molecular markers.

## MATERIALS AND METHODS

### Plant material

In this study 54 genotypes were used belonging to 17 *Coffea arabica* L. and one *Coffea canephora* Pierre ex A. Froehner accessions (Table 1) from the coffee germplasm collection of the coffee estate Santa Teresa (47°4'26" S, 88°5'20" W; 1513-1750 m a.s.l.), located in Villa Rica (89% RH annual average; 23.4 °C maximum, 15.8 °C minimum annual average temperature; 1978 mm annual precipitation), Provincia Oxapampa (Pasco Region), Peru. The origin of the genotypes is unknown, nevertheless, they were grouped in accessions according to the similarity of plant morphology and the cultivar name that was said to the farmer when acquired the seeds. Furthermore, Santa Teresa is a main producer and supplier of coffee botanical seeds to many farms in Peru. Besides that, their germplasm is included in the denomination of origin for Villa Rica coffee.

### DNA extraction

Genomic DNA was extracted from 0.1 g young and healthy leaves tissue of unstressed coffee plants. For this purpose, the protocol of Doyle and Doyle (1990) was used with the CTAB buffer (hexadecyltrimethylammonium bromide 3%, 2.1 mM NaCl, 250 mM Tris HCl pH 8, 50 mM EDTA pH 8) and some modifications (Palomino et al., 2014). Each

**Table 1. List of coffee accessions from the Santa Teresa estate of Villa Rica, Peru, with the corresponding code and species.**

Accession	Code	Genotype	Species	Accession	Code	Genotype	Species
1	JBrack	JBrack1	<i>Coffea arabica</i> L.	10	BorR	BorR1	<i>C. arabica</i>
		JBrack2	<i>C. arabica</i>			BorR2	<i>C. arabica</i>
		JBrack3	<i>C. arabica</i>			BorR3	<i>C. arabica</i>
2	Cat	Cat1	<i>C. arabica</i>	11	CatR	CatR1	<i>C. arabica</i>
		Cat2	<i>C. arabica</i>			CatR2	<i>C. arabica</i>
		Cat3	<i>C. arabica</i>			CatR3	<i>C. arabica</i>
3	Marg	Marg1	<i>C. arabica</i>	12	GrC	GrC1	<i>C. arabica</i>
		Marg2	<i>C. arabica</i>			GrC2	<i>C. arabica</i>
		Marg3	<i>C. arabica</i>			GrC3	<i>C. arabica</i>
4	Pac	Pac1	<i>C. arabica</i>	13	Caui	Caui1	<i>C. arabica</i>
		Pac2	<i>C. arabica</i>			Caui2	<i>C. arabica</i>
		Pac3	<i>C. arabica</i>			Caui3	<i>C. arabica</i>
5	BorA	BorA1	<i>C. arabica</i>	14	TypA	TypA1	<i>C. arabica</i>
		BorA2	<i>C. arabica</i>			TypA2	<i>C. arabica</i>
		BorA3	<i>C. arabica</i>			TypA3	<i>C. arabica</i>
6	Visar	Visar1	<i>C. arabica</i>	15	Lim	Lim1	<i>C. arabica</i>
		Visar2	<i>C. arabica</i>			Lim2	<i>C. arabica</i>
		Visar3	<i>C. arabica</i>			Lim3	<i>C. arabica</i>
7	Pch	Pch1	<i>C. arabica</i>	16	Rob	Rob1	<i>C. canephora</i> Pierre ex A. Froehner
		Pch2	<i>C. arabica</i>			Rob2	<i>C. canephora</i>
		Pch3	<i>C. arabica</i>			Rob3	<i>C. canephora</i>
8	CatA	CatA1	<i>C. arabica</i>	17	PchA	PchA1	<i>C. arabica</i>
		CatA2	<i>C. arabica</i>			PchA2	<i>C. arabica</i>
		CatA3	<i>C. arabica</i>			PchA3	<i>C. arabica</i>
9	GeiR	GeiR1	<i>C. arabica</i>	18	GeiA	GeiA1	<i>C. arabica</i>
		GeiR2	<i>C. arabica</i>			GeiA2	<i>C. arabica</i>
		GeiR3	<i>C. arabica</i>			GeiA3	<i>C. arabica</i>

sample was then digested with RNase (10 mg mL<sup>-1</sup>) to eliminate the remaining RNA. The concentration and purity were determined by measuring absorbance at 260 nm and the ratio 260/280 nm, respectively. The DNA integrity was determined by electrophoresis in 1% agarose gels using the TBE buffer solution (0.9 M Tris base, 0.9 M boric acid, 0.5 M EDTA, sterile milli-Q water), dyed with a nucleic acid gel stain (GelRed, Biotium, Fremont, California, USA) and visualized under UV light.

### Microsatellite markers

The PCR amplification was performed using 18 pairs of microsatellite primers (Table 2). The reaction mixture contained 60 ng DNA in 10 µL final volume. The master mix and the PCR amplification profiles were similar to those reported for each primer pair (Combes et al., 2000; Baruah et al., 2003; Silvestrini et al., 2007; Teressa et al., 2010; Geleta et al., 2012; Romero et al., 2014) but in some pairs the alignment temperatures were modified (Table 2). The amplification products were separated in 6% denaturing polyacrylamide gel and visualized with silver staining. The size of DNA fragments was estimated using a 100 bp marker.

### SRAP markers

The PCR amplification was obtained with 14 combinations of 12 primers (5 forward and 7 reverse) (Table 3) which were selected from Mishra et al. (2014) to show the highest percentages of polymorphism. The PCR amplification was carried out in 10 µL containing 60 ng DNA, 3 mM each primer, 2.5 mM MgCl<sub>2</sub>, 0.2 mM dNTPs, 1 U Taq DNA polymerase and 1X buffer solution. The amplification profile was set according to Mishra et al. (2014). The amplification products were separated by electrophoresis on 2.5% agarose gels and visualized under UV light. A DNA marker (50 to 1000 bp) was used to estimate the size of the fragments.

**Table 2. List of microsatellite markers, with the corresponding sequences of the primers, annealing temperatures in PCR and bibliographic source.**

Code of primer	Sequence of primers: Forward (F) and Reverse (R)	ATB ATU		Bibliographic source of primers
		°C		
M24	F: GGCTCGAGATATCTGTTTAG R: TTTAATGGGCATAGGGTCC	55	55	Combes et al., 2000
CMA199	F: CATGCCATCATCAATCCAT R: CTAGCTAGCTGGATCAGTACCC	50	55	Teressa et al., 2010
AY2434	F: CGCAAATGTTTATGTCAATC R: GCAACTTATGAGCCTAATCC	50	55	Teressa et al., 2010
Sat207	F: GAAGCCGTTTCAAGCC R: CAATCTCTTTCCGATGCTCT	51	56	Geleta et al., 2012
SSR03	F: GGACAAAACACCGCCAAAATA R: AGCGAGACAGAGGAAGGGAATATT	56	56	Geleta et al., 2012
CM5	F: GTAACCACCACCTCCTCTGC R: TGGAGGTAACGGAAGCTCTG	54	54	Geleta et al., 2012
E12-3CTG	F: TGCTTAGGCACTTGATATAGGA R: CACGTGCAAGTCACATACTTTA	60	65	Silvestrini et al., 2007
471	F: TTACCTCCCGCCAGAC R: CAGGAGACCAAGACCTTAGCA	54	58	Geleta et al., 2012
CMA055	F: TTGAGCAAAAACCCATTCC R: TAAACCCAAAAAGACCACAA	50	58	Teressa et al., 2010
CaM03	F: CGCGCTTGCTCCCTCTGTCTCT R: TGGGGGAGGGGCGGTGTT	62	64	Geleta et al., 2012
SSR09	F: TTGGCTTTTGTCCCTCCTTCTCTG R: AGCCATTTCCCTCTCATCTTCAAG	56	55	Geleta et al., 2012
CMA198	F: AGCAACTCCAGTCCTCAGGT R: TGGAAAGCCCGCATATAGTTT	50	56	Teressa et al., 2010
CM8	F: GCCAATTGTGCAAAGTGCT R: ATTCATGGGGCCTTTGTCTT	60	57	Baruah et al., 2003
AY 2449	F: CGAAAATATGCTGCCATTG R: CCGAACCATAAGGTGTGAC	50	60	Teressa et al., 2010
CaM16	F: AAGGCAGCTGAAGCGGGACAAA R: TGGGGAGAGCTGCAGTTGGAGG	62	62	Geleta et al., 2012
CM16	F: TGGGGAAAAGAAGGATATAGACAAGAG R: GAGGGGGGCTAAGGGAATAACATA	55	62	Baruah et al., 2003
SFcan022c	F: TTACTTTTGTTCCTCGTTCC R: CTTTTGGATATTCCGACTTG	53	50	Romero et al., 2014
05RM1	F: GTAGCCCATGTTGTGTGC R: CCAAACGGCTGAGTAACTTC	53	61	Romero et al., 2014

ATB: Annealing temperature in the bibliographic source; ATU: Annealing temperature used.

**Table 3. SRAP primers used, selected from Mishra et al. (2014).**

Forward primers		Reverse primers	
Code	Sequence	Code	Sequence
Me1	5'-TGAGTCCAAACCGGATA-3'	Em3	5'-GACTGCGTACGAATTGAC-3'
Me2	5'-TGAGTCCAAACCGGAGC-3'	Em4	5'-GACTGCGTACGAATTTGA-3'
Me3	5'-TGAGTCCAAACCGGAAT-3'	Em5	5'-GACTGCGTACGAATTAAC-3'
Me6	5'-TGAGTCCAAACCGGACA-3'	Em6	5'-GACTGCGTACGAATTGCA-3'
Me10	5'-TGAGTCCAAACCGGAAA-3'	Em11	5'-GACTGCGTACGAATTCTA-3'
		Em12	5'-GACTGCGTACGAATTC-3'
		Em13	5'-GACTGCGTACGAATTCTG-3'

## Data analysis

Amplicons from both microsatellite and SRAP markers were evaluated in the same way, by assigning 1 as presence and 0 as absence (Motta et al., 2014), because the microsatellite analysis of tetraploid genotypes like *C. arabica* does not allow to distinguish between different combinations of diallelic and triallelic genotypes in electrophoresis gels. Therefore, electrophoresis bands were considered as dominant markers in both marker systems (microsatellites and SRAPs). Hence, the genetic diversity and population structure could be calculated using both molecular marker systems together. In addition, the reproducibility of molecular markers was verified by repeated amplification.

Polymorphic information index (PIC), effective multiplex ratio (EMR) and marker index (MI) (Chesnokov and Artemyeva, 2015) were calculated to assess the molecular marker system.

In the genetic diversity analysis of coffee germplasm, the following descriptive parameters of the population were calculated: percentage of polymorphic loci, effective number of alleles ( $N_e$ ), gene diversity ( $h$ ) (Nei, 1973) and Shannon information index ( $I$ ) using the POPGENE software Version 1.32 (Yeh et al., 1999).

The population structure was analyzed by the coefficients of genetic differentiation ( $G_{ST}$ ) and the gene flow ( $N_m$ ), performed using POPGENE version 1.32 (Yeh et al., 1999). The composition of the variability was determined by analysis of molecular variance (AMOVA), using the Arlequin software view. 3.5 (Excoffier and Lischer, 2010). In addition, Nei's genetic distance matrix (Nei, 1978) of accessions was calculated with the POPGENE version 1.32 (Yeh et al., 1999). The grouping analysis with the unweighted pair group method with arithmetic mean (UPGMA) algorithm and the principal coordinate analysis were performed using DARWIN version 6.0.021 (Perrier and Jacquemoud-Collet, 2006). The population structure was corroborated with the Bayesian analysis through a bar chart, which was built by STRUCTURE software version 2.2 (Pritchard Lab, Department of Genetics, Stanford University, Stanford, California, USA). This analysis uses a Bayesian model that determines the groups of genetically similar genotypes (Pritchard et al., 2000). The analysis was performed with  $K$  values from one to 15, running 15 repetitions to each  $K$ , for which the mixing model (admixture) was applied with 400 000 iterations of the Markov Monte Carlo Chain (MCMC) and a period of 1 000 000 burn-in. The appropriate  $K$ -value was determined using the online program STRUCTURE HARVESTER web v0.6.94, which perform the Evanno method (Earl and vonHoldt, 2012).

## RESULTS

### Molecular marker assessment

Using 18 pairs of microsatellite marker primers, 100 bands were generated from the whole coffee germplasm collection, 97 of them were polymorphic (97% of the total); moreover, the means of PIC, ERM and MI were  $0.168 \pm 0.09$ ,  $5.285 \pm 2.72$  and  $0.909 \pm 0.65$ , respectively (Table 4). Table 4 shows that the largest number of polymorphic bands per primer pair was 14, which was obtained by CaM03, the smallest was two and was obtained by E12-3CTG. All the SSR markers had polymorphic bands.

The 14 combinations of SRAP primers generated 245 bands, 175 of which were polymorphic, representing 71.43% of the polymorphism; moreover, the means of the PIC, EMR, MI were  $0.112 \pm 0.04$ ,  $9.133 \pm 2.77$  and  $1.079 \pm 0.62$ , respectively (Table 4). In addition, all bands showed high quality and reproducibility (Figure 1). The largest number of polymorphic bands per primer pairs was 17 and was obtained with the Me6-Em5 combination, the lowest was six and was obtained with the Me2-Em6 combination (Table 4). The highest proportion of polymorphic bands was 91.67% which was obtained with the Me3-Em12 and Me1-Em6 combinations (Table 4).

### Genetic diversity analysis

Genetic diversity analysis of coffee accessions was carried out using both marker systems together. Nei (1973) genetic diversity index, Shannon index, effective number of alleles and percentage of polymorphic loci were calculated for each accession and for the total population (Table 5). Table 5 also shows that the PchA accession, belonging to *C. arabica* species, shows the highest values of the genetic diversity index ( $h = 0.0967$ ), Shannon index ( $I = 0.137$ ) and effective number of alleles ( $N_e = 1.1834$ ). The Lim accession on the other hand has the lowest values for such estimators ( $h = 0.006$ ,  $I = 0.0087$ , and  $N_e = 1.0108$ ), while the Rob accession, belonging to the *C. canephora* species, showed the highest values for number and percentage of polymorphic loci (74 and 21.45, respectively).

**Table 4. Results of the marker systems analysis in the coffee germplasm from the Santa Teresa estate of Villa Rica, Peru.**

Primers pairs for microsatellite/SRAP	MB	P	T	Range of amplicons bp	APB %	PIC	EMR	IM
M24	0	6	6	132-212	100.00	0.2973	6.0000	1.7840
CMA199	0	6	6	137-163	100.00	0.2620	6.0000	1.5720
AY2434	0	4	4	186-200	100.00	0.0881	4.0000	0.3525
Sat207	0	7	7	78-98	100.00	0.0990	7.0000	0.6927
SSR03	0	7	7	161-189	100.00	0.1564	7.0000	1.0947
CM5	0	6	6	188-270	100.00	0.1049	6.0000	0.6296
E12-3CTG	1	2	3	149-187	66.67	0.0588	1.3333	0.0783
471	0	6	6	300-362	100.00	0.4214	6.0000	2.5284
CMA055	0	6	6	89-109	100.00	0.2651	6.0000	1.5905
CaM03	0	14	14	170-200	100.00	0.1101	14.0000	1.5412
SSR09	0	4	4	129-153	100.00	0.1702	4.0000	0.6808
CMA198	0	5	5	195-251	100.00	0.0778	5.0000	0.3889
CM8	0	4	4	172-222	100.00	0.1975	4.0000	0.7901
AY2449	0	4	4	296-333	100.00	0.0881	4.0000	0.3525
CaM16	0	3	3	211-222	100.00	0.1738	3.0000	0.5213
CM16	2	3	5	106-128	60.00	0.1576	1.8000	0.2836
SFcan022c	0	4	4	593-630	100.00	0.1408	4.0000	0.5631
05RM1	0	6	6	792-891	100.00	0.1527	6.0000	0.9163
Me1-Em12	4	7	11	80-2400	63.64	0.0834	4.4545	0.3714
Me2-Em4	7	10	17	190-2900	58.82	0.0707	5.8824	0.4158
Me3-Em11	8	16	24	110-475	66.67	0.0789	10.6667	0.8413
Me6-Em5	7	17	24	110-475	70.83	0.1279	12.0417	1.5404
Me10-Em13	8	11	19	80-490	57.89	0.0862	6.3684	0.5489
Me3-Em4	10	16	26	90-1500	61.54	0.0545	9.8462	0.5364
Me2-Em12	5	16	21	180-2900	76.19	0.1792	12.1905	2.1846
Me1-Em5	8	13	21	130-340	61.90	0.0938	8.0476	0.7548
Me3-Em3	4	14	18	400-3000	77.78	0.0958	10.8889	1.0427
Me3-Em12	1	11	12	100-2000	91.67	0.2121	10.0833	2.1387
Me1-Em6	1	11	12	390-2200	91.67	0.1296	10.0833	1.3065
Me1-Em4	3	13	16	170-490	81.25	0.1399	10.5625	1.4774
Me2-Em3	2	14	16	150-2000	87.50	0.1229	12.2500	1.5055
Me2-Em6	2	6	8	250-3000	75.00	0.0976	4.5000	0.4390
Total/Average (microsatellite)	3	97	100		97.00	0.1679 (0.09)	5.2852 (2.72)	0.9089 (0.65)
Total/Average (SRAP)	70	175	245		71.43	0.1123 (0.04)	9.1333 (2.77)	1.0788 (0.62)

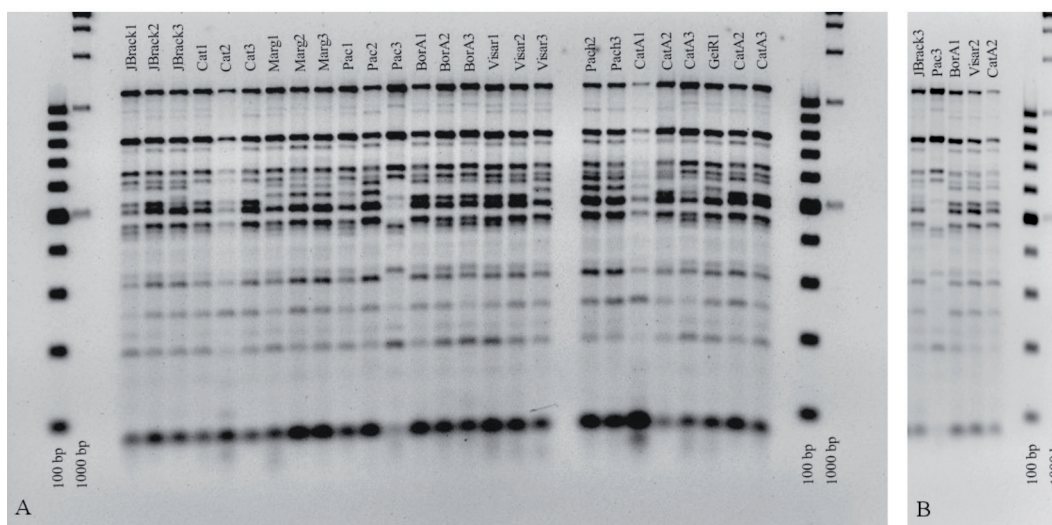
MB: Number of monomorphic bands; PB: number of polymorphic bands; TB: total number of bands; APB: average of polymorphic bands; PIC: polymorphic index content; EMR: effective multiplex ratio; IM: marker index; SRAP: sequence-related amplified polymorphism. The values in parentheses are standard deviation.

Table 6 reveals a slight decrease in genetic diversity values ( $h = 0.096$ ;  $I = 0.1614$ ;  $N_e = 1.1428$ ; 178 polymorphic loci, representing 51.59% of the total bands) when the *C. canephora* accession was excluded. In addition, the calculation of genetic diversity estimators in *C. arabica* accessions for each molecular marker systems separately showed that SSR markers had higher average values than SRAP markers (Table 6).

### Analysis of the population structure

Analysis of the population structure was carried out using only the *C. arabica* accessions. The genetic differentiation  $G_{ST}$  showed high values ( $G_{STSRAP} = 0.6033$ ,  $G_{STSSR} = 0.7568$  and  $G_{STSRAP+SSR} = 0.6584$ ); while gene flow ( $N_m$ ) had low values ( $N_{mSRAP} = 0.3288$ ,  $N_{mSSR} = 0.1606$  and  $N_{mSRAP+SSR} = 0.2594$ ) (Table 6). In addition, AMOVA analysis allowed to understand the genetic potential in term of accessions variability (Table 7): the variation among and within accessions was 43.05% and 56.95%, respectively, while the fixation index ( $F_{ST}$ ) was 0.43045 (p-value < 0.0001).

**Figure 1. Amplification pattern generated by sequence-related amplified polymorphism (SRAP) for 26 genotypes of coffee plant (A) and five samples amplified in another PCR reaction (B).**



**Table 5. List of accessions of *Coffea arabica* with the respective estimators of genetic diversity.**

Accession	Code	Number of genotypes	$N_e$	h	I	Number of PL	PPL
1	JBrack	3	1.0679 (0.2779)	0.0367 (0.1236)	0.0525 (0.1756)	29	8.41
2	Cat	3	1.0283 (0.1516)	0.0157 (0.0812)	0.0227 (0.1165)	13	3.77
3	Marg	3	1.0824 (0.2520)	0.0452 (0.1340)	0.0653 (0.1915)	37	10.72
4	Pac	3	1.1298 (0.3068)	0.0711 (0.1627)	0.1027 (0.2321)	58	16.81
5	BorA	3	1.0600 (0.2097)	0.0342 (0.1142)	0.0503 (0.1661)	30	8.70
6	Visar	3	1.0409 (0.1865)	0.0219 (0.0978)	0.0312 (0.1386)	17	4.93
7	Pch	3	1.0117 (0.0884)	0.0072 (0.0514)	0.0109 (0.0767)	7	2.03
8	CatA	3	1.0253 (0.1376)	0.0146 (0.0760)	0.0215 (0.1107)	13	3.77
9	GeiR	3	1.0671 (0.2247)	0.0376 (0.1212)	0.0549 (0.1746)	32	9.28
10	BorR	3	1.0655 (0.2332)	0.0349 (0.1220)	0.0498 (0.1726)	27	7.83
11	CatR	3	1.0458 (0.1910)	0.0253 (0.1022)	0.0367 (0.1465)	21	6.09
12	GrC	3	1.0284 (0.1312)	0.0178 (0.0781)	0.0272 (0.1179)	18	5.22
13	Caui	3	1.0545 (0.1885)	0.0327 (0.1076)	0.0491 (0.1592)	31	8.99
14	TypA	3	1.0984 (0.2707)	0.0544 (0.1444)	0.0788 (0.2068)	45	13.04
15	Lim	3	1.0108 (0.0942)	0.0060 (0.0506)	0.0087 (0.0727)	5	1.45
16	Rob*	3	1.1390 (0.2922)	0.0812 (0.1615)	0.1206 (0.2358)	74	21.45
17	PchA	3	1.1834 (0.3658)	0.0967 (0.1897)	0.1370 (0.2671)	73	21.16
18	GeiA	3	1.0188 (0.1224)	0.0105 (0.0663)	0.0154 (0.0956)	9	2.61

$N_e$ : Number of effective alleles; h: Nei (1973) diversity index; I: Shannon index; PL: polymorphic loci; PPL: percentage of polymorphic loci. The values in parentheses are standard deviation.

\**Coffea canephora*.

**Table 6. Genetic diversity and population structure indices for the two molecular marker systems applied to the accessions of *Coffea arabica*.**

Marker system	Samples	PL	PPL (%)	$N_e$	h	I	$G_{ST}$	$N_m$
SRAP	51	116	47.35	1.1292 (0.2306)	0.0874 (0.1351)	0.1465 (0.2037)	0.6033	0.3288
SSR	51	62	62.00	1.1759 (0.2500)	0.1189 (0.1486)	0.1979 (0.2184)	0.7568	0.1606
SRAP + SSR	51	178	51.59	1.1428 (0.2370)	0.0966 (0.1394)	0.1614 (0.2090)	0.6584	0.2594

PL: Polymorphic loci; PPL: percentage of polymorphic loci;  $N_e$ : Number of effective alleles; h: Nei (1973) diversity index; I: Shannon index;  $G_{ST}$ : genetic differentiation;  $N_m$ : gene flow.

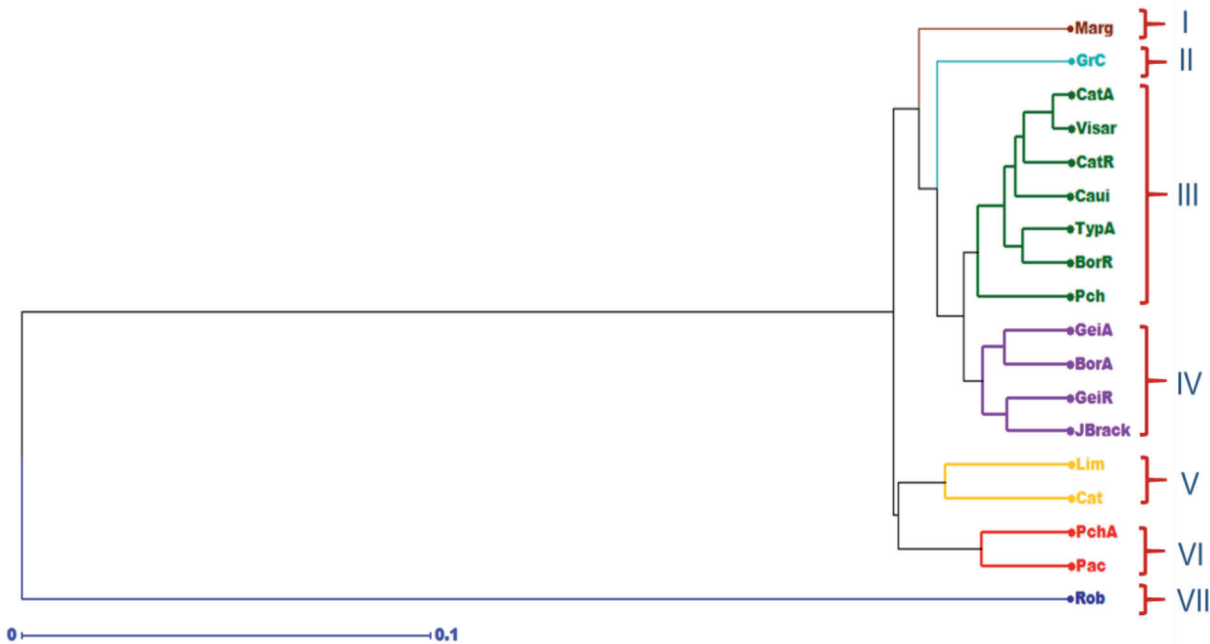
The values in parentheses are standard deviation.

**Table 7. Analysis of molecular variance (AMOVA) of the *Coffea arabica* germplasm from the Santa Teresa estate of Villa Rica, Peru.**

Variation source	Degrees of freedom	Sum of squares	Components of variance	Percentage of variation
Between accessions	16	476.137	6.88358 Va	43.05
Within accessions	34	309.667	9.10784 Vb	56.95
Total	50	785.804	15.99142	

The genetic distance matrix was then used to build an UPGMA dendrogram (Figure 2) to assess the relationships among 18 accessions from the coffee collection. There are no duplicate accessions in the dendrogram, therefore each accession is different from the other. In addition, at 0.1 of distance, two groups appear well separated and structured: the first group contains only the *C. canephora* accession while the second contains all *C. arabica* accessions (Arabica). Furthermore, the Arabica coffee group can be further divided into six subgroups: the first two are composed of a single accession (Marg and GrC, respectively); the third is the largest and includes CatA, Visar, CatR, Caui, TypA and Pch accessions; the fourth consists of GeiA, BorA, GeiR and JBrack; the fifth is made up of Lim and Cat, while the sixth is composed by PchA and Pac. The principal coordinate analysis was performed using the same distance matrix (Figure 3), with the x-axis corresponding to the coordinate one and the y-axis corresponding to the coordinate two. In Figure 3, the Rob accession of *C. canephora* is separated from the *C. arabica* accessions. Likewise, the analysis of the genetic structure performed by the Bayesian method on the 54 coffee tree genotypes from the collection was carried out using the mixing model. In the determination of the number of groups with the delta K of Evanno method, a first peak was generated at K = 2, therefore two groups can be identified in the collection: the first with the *C. canephora* genotype and the second containing the *C. arabica* accessions. However, the delta K method also generated a second peak at K = 7, suggesting the existence of the same six subgroups of *C. arabica* generated by the UPGMA dendrogram (Figures 2 and 4).

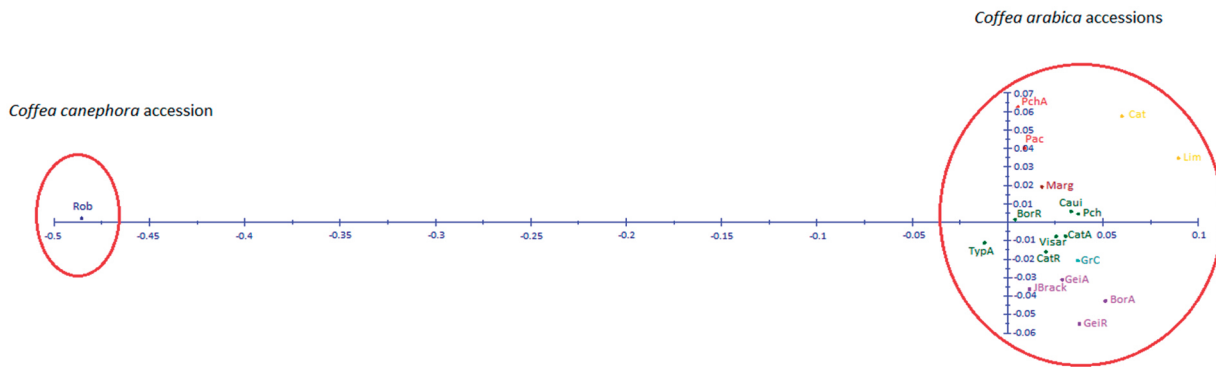
**Figure 2. Dendrogram of coffee accessions from the Santa Teresa estate of Villa Rica, Peru, generated using the unweighted pair group method with arithmetic mean (UPGMA) algorithm.**



Groups are numbered in Roman numerals, from I to VI are *Coffea arabica* accessions while VII is *C. canephora*.

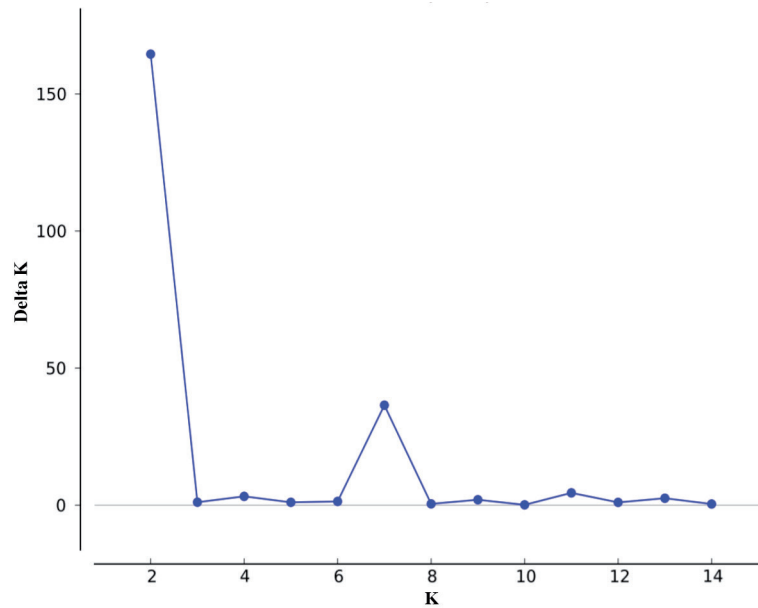


**Figure 3.** Principal coordinates analysis carried out on the germplasm collection of Santa Teresa estate of Villa Rica, Peru.



The x-axis corresponds to component 1 and the y-axis to component 2. Rob is the *Coffea canephora* accession.

**Figure 4.** Results of delta K Evanno method for the population structure analysis of the germplasm collection of Santa Teresa estate of Villa Rica, Peru (the highest peak is shown at K = 2 with a second peak at K = 7).



## DISCUSSION

The evaluation of the microsatellite and SRAP markers systems in terms of percentage of polymorphism, PIC and MI allowed their combined use to analyze the population structure and genetic diversity of the coffee germplasm collection under study. Firstly, the percentages of polymorphism in both systems were high, even when only *C. arabica* accessions were evaluated (47.35% in SRAP and 62% in SSR) (Table 6). These results are congruent with the studies by Teresa et al. (2010) and Jingade et al. (2019) where SSR and SRAP markers were used, respectively. In addition, the analysis revealed a PIC index for SSR (PIC = 0.1679) higher than for SRAP (PIC = 0.1123), while, on the other hand, the MI index, which measures the efficiency of the molecular marker system, was lower for microsatellites (MI = 0.9089) than for SRAP (MI = 1.0788) (Table 4). Therefore, based on the IM, the SRAP would be a more efficient tool to determine the polymorphism, even if its PIC average value is low. Moreover, the electrophoresis in agarose gel used in SRAP markers is faster and less laborious than the electrophoresis in polyacrylamide gel for microsatellites. On the other hand,

Casadevall et al. (2011), using microsatellites and SRAP markers to study the genetic diversity of *Cynara cardunculus*, recommended to use microsatellite markers for their easy visualization and codominance information. The fact that in this study as well as in our research (Table 6) the estimators of genetic diversity obtained by SRAP appear lower than those obtained by microsatellite might be due to the characteristics of the marker system. The SRAP markers are focused on chromosome regions rich in genes and the corresponding polymorphic bands are based on introns and exons, whilst, according to Abdurakhmonov (2016), the microsatellites markers are mainly linked to non-coding genomic regions like intergenic spaces and introns. Therefore, considering the polyploidy of *C. arabica* and the results of our analysis of marker systems, we believe that SSR and SRAP marker systems should both be used to estimate the genetic diversity and structure population of *C. arabica* Peruvian germplasm.

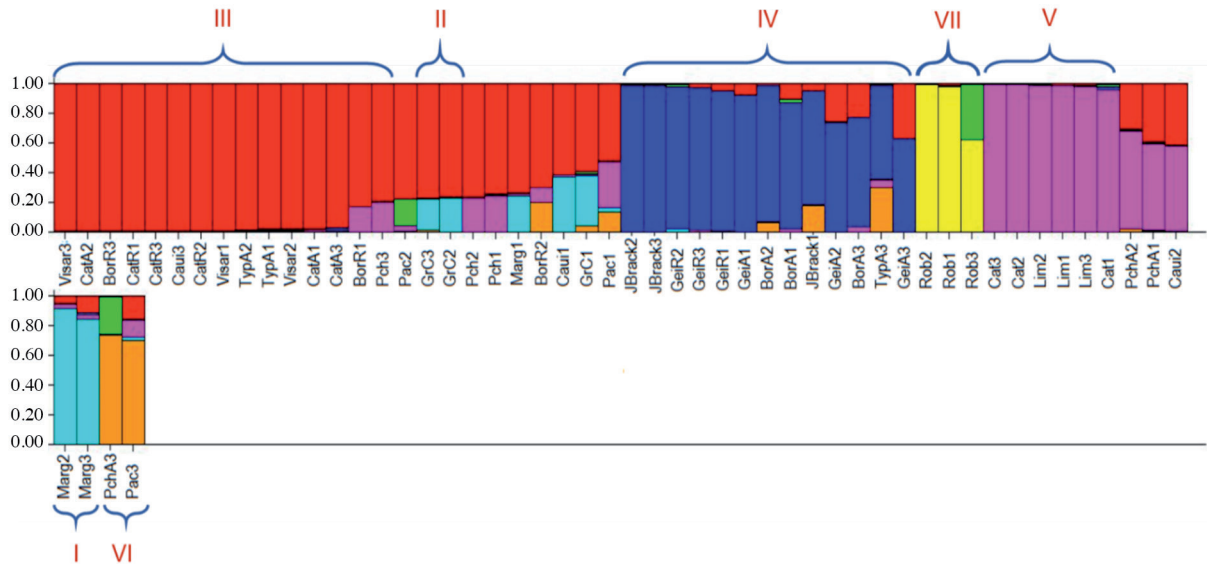
To study the genetic diversity of the coffee accessions, only one database was generated by screening the gel electrophoretic profiles of both markers system. The percentage of polymorphic bands obtained was high (51.59%) (Table 6) and is within the range of polymorphism reported in previous studies of *C. arabica* genetic diversity (Teresa et al., 2010; Geleta et al., 2012; Al-Murish et al., 2013; Mishra et al., 2014; Romero et al., 2014; Jingade et al., 2019). Furthermore, this would indicate the existence of a high molecular variability in the germplasm evaluated.

On the other hand, other genetic diversity estimators obtained in the germplasm of *C. arabica* were low. The index of genetic diversity value obtained (Table 6) was lower than the one reported by Geleta et al. (2012), who analyzed a bigger population from the main coffee producer areas of Nicaragua. The Shannon index ( $I = 0.1614$ ) was lower to the one reported by Silvestrini et al. (2007) in a collection from Ethiopia, Eritrea, Yemen and Brazil ( $I = 0.21$ ) and the one by Aga et al. (2003) ( $I = 0.3$ ) in Arabic coffee from Ethiopian forests. However, Silvestrini et al. (2007) and Scalabrin et al. (2020) both suggested that the *C. arabica* diversity is very low as compared to other crops because of the recent hybrid origin of the species and of the way the world commercial cultivars have been generated from a very low number of genotypes. In addition, the preferred autogamic reproductive fashion of the species allows only a 10% of cross fertilization. On other side, the low genetic diversity index obtained for *C. arabica* is in contrast to the high index obtained for *C. canephora* (Table 6), similarly to Silvestrini et al. (2007) and Motta et al. (2014).

The high index of genetic differentiation ( $G_{ST} = 0.6584$ ), similar to what obtained by Silvestrini et al. (2007), might be mainly due to the autogamic reproduction of *C. arabica*. Therefore, the different accessions of the germplasm collection might be all different from each other and, even if they grow in a same farm, their genetic flux was low ( $N_m = 0.2594$ ). Furthermore, these accessions might not be very different from the cultivars from which they were generated, after being introduced to Peru. Furthermore, the genetic diversity index higher than zero for each accession would indicate that the level of genetic diversity among the Peruvian cultivars is relatively high, similarly to what reported by Geleta et al. (2012) in Nicaragua. In terms of the origin of cultivars grown in Peru, there are written reports only limited to the Spanish viceroyalty and the beginning of the republican period. Furthermore, it cannot be ruled out that improved plants could have also been introduced from Central America, Brazil and Colombia, which would present, among others, genes introgressed mainly from the Timor hybrid. It is also possible that in some areas of Peru there are still descendants from a group of genotypes arrived in Peru from Ethiopia and Eritrea thanks to a FAO expedition of 1964-1965, according to Silvestrini et al. (2007). Therefore, in the different coffee producing areas in Peru there might be germplasm, whose genetic diversity is not yet evaluated, that could be used in genetic improvement programs.

The AMOVA (Table 7), that assigns the most variability to the inter-accessions source, allowed the structuring of the coffee germplasm into different accessions. Furthermore, the high  $F_{ST}$  value reinforced the hypothesis of low gene flux between accessions. Besides, the grouping of accessions generated from the UPGMA dendrogram and principal coordinates analysis was congruent with the Bayesian analysis. In this analysis, the graph generated by STRUCTURE with  $K = 2$  has clearly revealed the expected genomic differences between *C. canephora* and *C. arabica*, as revealed also by UPGMA dendrogram and principal coordinates analyses. When the Bayesian analysis was performed with  $K = 7$ , it allowed to clearly group the Arabica accessions into six clusters that correspond to those observed in the UPGMA dendrogram (Figures 2 and 5). This clustering of *C. arabica* accessions might be due to the relationships existing among the accessions within each cluster, because they probably came from the same cultivar or cultivars that are genetically close. Nevertheless, some accessions like PchA ( $h = 0.0967$ ) have a genetic diversity higher than the remaining accessions. This relatively high diversity would be explained by the different origins of genotypes, which could have arrived to the farm at different times, but that were grouped together because of their morphological similarity.

**Figure 5. Population structure of 54 genotypes from the collection of coffee trees of Santa Teresa estate in Villa Rica, Peru, generated using the STRUCTURE software for K = 7.**



Each color represents a subpopulation, while the Roman numerals from I to VII identify the groups of accessions generated with the UPGMA dendrogram.

Nevertheless, the high differentiation and low gene flow between the accessions of the present study contrast with Palomino et al. (2014), who, using only one molecular marker system, proposed that in Villa Rica there is a mix of cultivars. The analysis based on the combination of SRAP and microsatellite molecular marker systems seemed able to differentiate the different cultivars, therefore might shed some light on the relationships existing among coffee cultivars grown in Peru.

## CONCLUSIONS

The high genetic differentiation between the coffee accessions in the collection, the genetic structuring within the *Coffea arabica* accessions and the low percentage of cross pollination would indicate that the cultivars, from which the accessions were originated, have been preserved over time. In addition, the genetic diversity that is maintained in the different coffee producing areas in Peru would be the associated with the bicentennial presence of coffee in Peru along with the reported and unreported introduction of improved cultivars from different origins.

## ACKNOWLEDGEMENTS

The authors would like to thank Hans Brack Frantzen manager of Fundo Santa Teresa for vegetal material and, the Proyecto PNIA 017-2015-INIA-PNIA for financial support.

## REFERENCES

- Abdurakhmonov, I.Y. (ed.) 2016. Introduction to microsatellites: Basics, trends and highlights. p. 3-16. In *Microsatellite markers*. IntechOpen. doi:<http://dx.doi.org/10.5772/66446>.
- Aerts, R., Berecha, G., Gijbels, P., Hundera, K., Van Glabeke, S., Vandepitte, K., et al. 2012. Genetic variation and risks of introgression in the wild *Coffea arabica* gene pool in south-western Ethiopian montane rainforests. *Evolutionary Applications* 6:243-252. doi:10.1111/j.1752-4571.2012.00285.x.
- Aga, E., Bryngelsson, T., Bekele, E., and 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(1):36-46.

- Alemayehu, D. 2017. Review on genetic diversity of coffee (*Coffea arabica* L.) in Ethiopia. *International Journal of Forestry and Horticulture* 3(2):18-27. doi:10.20431/2454-9487.0302003.
- Al-Murish, M.T., Elshafei, A.A., Al-Doss, A.A., and Barakat, M.N. 2013. Genetic diversity of coffee (*Coffea arabica* L.) in Yemen via SRAP, TRAP and SSR markers. *Journal of Food Agriculture and Environment* 11(2):411-416.
- Baruah, A., Naik, V., Hendre, P., Rajkumar, R., Rajendrakumar, P., and Aggarwal, R. 2003. Isolation and characterization of nine microsatellite markers from *Coffea arabica* L., showing wide cross-species amplifications. *Molecular Ecology Resources* 3(4):647-650.
- Casadevall, R., Martin, E., and Cravero, V. 2011. Simple sequence repeat (SSR) vs. sequence-related amplified polymorphism (SRAP) markers for *Cynara cardunculus* characterization. *Spanish Journal of Agricultural Research* 9(2):453-459.
- Chesnokov, Y.V., and Artemyeva, A.M. 2015. Evaluation of the measure of polymorphism information of genetic diversity. *Agricultural Biology* 50(5):571-578.
- Combes, M.C., Andrzejewski, S., Anthony, F., Bertrand, B., Rovelli, P., Graziosi, G., et al. 2000. Characterization of microsatellite loci in *Coffea arabica* and related coffee species. *Molecular Ecology* 9(8):1178-1180.
- Doyle, J., and Doyle, L. 1990. Isolation of DNA from small amounts of plant tissues. *Focus* 12:13-15.
- Earl, D.A., and vonHoldt, B.M. 2012. Structure harvester: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* 4(2):359-361.
- Excoffier, L., and Lischer, H.E.L. 2010. Arlequin suite ver 3.5: A new series of programs to perform population genetics analysis under Linux and Windows. *Molecular Ecology Resources* 10:564-567.
- Ferrao, L.F.V., Caixeta, E.T., Pena, G., Zambolim, E.M., Cruz, C.D., Zambolim, L., et al. 2015. New EST-SSR markers of *Coffea arabica*: transferability and application to studies of molecular characterization and genetic mapping. *Molecular Breeding* 35:31. doi:10.1007/s11032-015-0247-z.
- Geleta, M., Herrera, I., Monzón, A., and Bryngelsson, T. 2012. Genetic diversity of arabica coffee (*Coffea arabica* L.) in Nicaragua as estimated by simple sequence repeat markers. *The Scientific World Journal* 2012:128-145. doi:10.100/2012/939820.
- Jingade, P., Huded, A.K., Kosaraju, B., and Mishra, M.K. 2019. Diversity genotyping of Indian coffee (*Coffea arabica* L.) germplasm accessions by using SRAP markers. *Journal of Crop Improvement* 33(3):327-345.
- Mishra, M., Nishani, S., Gowda, M., Padmajyothi, D., Suresh, N., Sreenath, H., et al. 2014. Genetic diversity among Ethiopian coffee (*Coffea arabica* L.) collections available in Indian gene bank using sequence related amplified polymorphism markers. *Plant Breeding and Seed Science* 70(1):29-40.
- Motta, L.B., Soares, T.C.B., Ferrao, M.A.G., Caixeta, E.T., Lorenzoni, R.M., and Neto, J.D.S. 2014. Molecular characterization of arabica and conilon coffee plants genotypes by SSR and ISSR markers. *Brazilian Archives of Biology and Technology* 57:728-735. doi:10.1590/S1516-8913201402071.
- Nei, M. 1973. Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America* 70(12):3321-3323.
- Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89:583-590.
- Palomino, C., López, C., Espejo, R., Mansilla, R., y Quispe, J. 2014. Evaluación de la diversidad genética del café (*Coffea arabica* L.) en Villa Rica (Perú). *Ecología Aplicada* 13(2):129-134.
- Perrier, X., and Jacquemoud-Collet, J.P. 2006. DARwin software, Montpellier, France. Available at <http://darwin.cirad.fr/> (accessed January 2020).
- Pestana, K.N., Capucho, A.S., Caixeta, E.T., de Almeida, D.P., Zambolim, E.M., Cruz, C.D., et al. 2015. Inheritance study and linkage mapping of resistance loci to *Hemileia vastatrix* in híbrido de Timor UFV 443-03. *Tree Genetics and Genomes* 11:72. doi:10.1007/s11295-015-0903-9.
- Pritchard, J.K., Stephens, M., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945-959.
- Romero, G., Alvarado, G., Ligarreto, G., and Galeano, N.F. 2010. Partial resistance to leaf rust (*Hemileia vastatrix*) in coffee (*Coffea arabica* L.): genetic analysis and molecular characterization of putative candidate genes. *Molecular Breeding* 25(4):685-697.
- Romero, G., Vásquez, L.M., Lashermes, P., and Herrera, J.C. 2014. Identification of a major QTL for adult plant resistance to coffee leaf rust (*Hemileia vastatrix*) in the natural Timor hybrid (*Coffea arabica* x *C. canephora*). *Plant Breeding* 133(1):121-129.
- Scalabrin, S., Toniutti, L., DiGasparo, G., Scaglione, D., Magris, G., Vidotto, M., et al. 2020. A single polyploidization event at the origin of the tetraploid genome of *Coffea arabica* is responsible for the extremely low genetic variation in wild and cultivated germplasm. *Scientific Reports* 10:4642. doi:10.1038/s41598-020-61216-7.
- Silvestrini, M., Junqueira, M., Favarin, A., Guerreiro-Filho, O., Maluf, M., Silvarolla, M., et al. 2007. Genetic diversity and structure of Ethiopian, Yemen and Brazilian *Coffea arabica* L. accessions using microsatellites markers. *Genetic Resources and Crop Evolution* 54(6):1367-1379.
- Teressa, A., Crouzillat, D., Petiard, V., and Brouhan P. 2010. Genetic diversity of Arabica coffee (*Coffea arabica* L.) collections. *Ethiopian Journal of Science and Technology* 1(1):63-79.
- Van Der Vossen, H., Bertrand, H., and Charrier, A. 2015. Next generation variety development for sustainable production of arabica coffee (*Coffea arabica* L.): a review. *Euphytica* 204:243-256.

- Vieira, E., Teixeira, E., Alkimim, E.R., Baião de Oliveira, A.C., Pereira, A.A., Sakiyama, N.S., et al. 2017. Population structure and genetic diversity of coffee progenies derived from Catuaí and Híbrido de Timor revealed by genome-wide SNP marker. *Tree Genetics and Genomes* 13:124. doi:0.1007/s11295-017-1208-y.
- Vieira, E., Von Pinho, É., Carvalho, M., Esselink, D., and Vosman, B. 2010. Development of microsatellite markers for identifying Brazilian *Coffea arabica* varieties. *Genetics and Molecular Biology* 33:507-514.
- Yan, L., Ogutu, C., Huang, L., Wang, X., Zhou, H., Lv, Y., et al. 2019. Genetic diversity and population structure of coffee germplasm collections in China revealed by ISSR markers. *Plant Molecular Biology Reporter* 37:204-213. doi:10.1007/s11105-019-01148-3.
- Yeh, F.C., Yang, R.C., and Boyle, T. 1999. POPGENE Microsoft Window-based freeware for population genetic analysis, Release 1.31. University of Alberta, Edmonton, Canada. Available at <https://sites.ualberta.ca/~fyeh/popgene.html> (accessed January 2020).