Genet Resour Crop Evol (2020) 67:1795–1806
https://doi.org/10.1007/s10722-020-00940-5

RESEARCH ARTICLE

Genetic diversity of coffee (*Coffea* spp.) in Mexico evaluated by using DArTseq and SNP markers

J. L. Spinoso-Castillo · E. Escamilla-Prado · V. H. Aguilar-Rincón · V. Morales Ramos · G. García de los Santos · P. Pérez-Rodríguez · T. Corona-Torres

Received: 24 October 2019 / Accepted: 6 April 2020 / Published online: 25 April 2020

Abstract The *Coffea* genus includes about 124 species, of which *Coffea arabica* L. and *Coffea canephora* Pierre ex A. Froehner are the most commercially important, supplying 63 and 37% of the world coffee production, respectively. *C. arabica* L. is associated with better cup quality, but presents serious problems with pests and diseases, potentiated by climate change. The development of new varieties of *C. arabica* L. requires the inclusion of genetic diversity, looking for incorporating new genes to tolerate or resist the most important pests and diseases, maintaining or improving yield and cup quality of the new materials. The technology of diversity arrays based on sequencing (DArTseq^{TM}) allows the rapid identification of the single nucleotide polymorphism (SNP) for a genomic approach of genetic diversity. The objectives of the present study were: (1) to evaluate the diversity and genetic structure of a central coffee collection of the Mexican National Bank of Coffee Germplasm; (2) assess the reproducibility and error rates of the markers and their broad representation in the genome and (3) to propose a new collection with representative and divergent promising genotypes for stablishing a coffee breeding program in Mexico. 87 coffee accessions of three different species of the *Coffea* genus were sequenced and 16,995 SNP markers were derived from 34,000 unique sequences obtained by the DArTseq^{TM} method. After removing the markers with more than 10% of missing data and the MAF $< 5\%$, we end up with 1739 polymorphic SNP markers for the analysis. A dendrogram was generated which divided the accessions into five divergent groups where promising genotypes were identified. In conclusion, there is genetic diversity among the accessions of *Coffea* spp. for use in future breeding programs.

Keywords *Coffea canephora* · *Coffea arabica* · *Coffea liberica* · High throughput genotyping · Molecular markers · Plant breeding

Introduction

The *Coffea* genus belongs to the Rubiaceae family that includes around 124 species, most of them are diploids

J. L. Spinoso-Castillo · V. H. Aguilar-Rincón ·
G. G. de los Santos · P. Pérez-Rodríguez ·
T. Corona-Torres (✉)
Colegio de Postgraduados, Campus Montecillo, Carretera
México-Texcoco Km 36.5, 56230 Texcoco,
Estado de México, Mexico
e-mail: tcoronat@gmail.com

E. Escamilla-Prado
Universidad Autónoma Chapingo, Centro Regional
Universitario Oriente, Carretera Huatusco-Xalapa Km 6,
94100 Huatusco, Veracruz, México

V. Morales Ramos
Colegio de Postgraduados Campus Córdoba, Carretera
Córdoba Veracruz Km 348, 94946 Amatlán de los Reyes,
Veracruz, México

(2n = 2x = 22). The only allotetraploid is *C. arabica* L., with 2n = 4x = 44 (Davis et al. 2011) which was originated from the natural cross between *Coffea eugenioides* S. Moore and *C. canephora* Pierre ex A. Froehner (Lashermes et al. 1999), and it is autogamous with approximately 10% cross-fertilization (Carvalho and Krug 1949). *C. arabica* and *C. canephora* are the most important agronomic species, which supplied 63 and 37% of the world coffee production in 2016, respectively (ICO 2017).

One of the main objectives of breeding programs is to create more productive cultivars, adapted to the local conditions of interest. Some limitations faced by coffee breeders are the long time (about 25 years) and the considerable resources needed to develop new varieties due to the perennial nature of these species (Moreno 2004). An important challenge is the reduced genetic variability available in commercial plantations (Moncada et al. 2016).

Coffee plantations in Mexico include the cultivars Typica, Bourbon, Caturra Rojo, Mundo Novo, Garnica and Caturra Amarillo, which are susceptible to coffee leaf rust (*Hemileia vastratix* Berk & Br.) (Escamilla et al. 2005; López-García et al. 2016). Plant breeding for resistance to this disease is the best long-term solution (Avelino et al. 2015). Thus, applying molecular markers is particularly desirable for *C. arabica* due to its narrow genetic base (Ferrão et al. 2015). Molecular information, when combined with phenotypic variables, allows the selection of superior genotypes and maximizes the selection gains (Sousa et al. 2017) as the breeders select more diverse germplasm and avoid the crossing of closely related accessions (Pailles et al. 2017). Molecular markers have shown that the genetic diversity of *C. arabica* is lower than *C. canephora* (Cubry et al. 2008; Lashermes et al. 2011; Ferrão et al. 2015).

There is a new highly informative and high-performance genome marker technology, called DArT genotyping by sequencing (DArTseq^{TM}). This technology uses the DArT markers platform and, combined with next-generation sequencing, allows for rapid identification of single nucleotide polymorphism (SNP) (Kilian et al. 2012; Cruz et al. 2013; Raman et al. 2014). 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 (Zhou et al. 2016). DArTseq^{TM} has been applied successfully in the evaluation of the genetic diversity of *Solanum lycopersicum* (Pailles et al. 2017), *Solanum tuberosum* (Berdugo-Cely et al. 2017), *Allium sativum* (Egea et al. 2017) and in the *Coffea* genus is reported with *C. canephora* (Garavito et al. 2016).

Since the Mexican Coffee Institute (INMECAFE) closed down in 1989, Mexico has implemented few coffee breeding programs, importing most of the leaf rust coffee resistant cultivars to face the sanitary crisis of 2012. The term “Central Collection” refers to the subset of accessions of a larger collection that includes, with minimum redundancy, the majority of the genetic diversity of a crop, a wild species or a group of species (Van Hintum et al. 2000). In this sense, this work tries to develop a “Core Collection” representative of the Central Collection that is conserved in the National Bank of Coffee Germplasm located in Huatusco, Veracruz, Mexico. The objectives of this study are: (1) to evaluate the diversity and genetic structure of the central coffee collection; (2) to assess the reproducibility and error rates of the markers and their broad representation in the genome and (3) to propose a new collection with representative and divergent promising genotypes for stablishing a coffee breeding program in Mexico.

Materials and methods

Plant material and DNA extraction

A total of 87 accessions of *Coffea* spp. (Table 1) from the National Bank of Coffee Germplasm located at 19° 10′ 27″ N and 96° 57′ 50″ W and 1345 masl, in Huatusco, Veracruz, Mexico were characterized by DArTseq^{TM} method and SNP markers. Six young and fully expanded leaves were collected from a single individual plant per accession and stored in a freezer at – 80 °C until use. Genomic DNA was extracted from previously freezed leaves by the CTAB method (cetyltrimethylammonium bromide) (Hoisington et al. 1994), with two additional chloroform washes for further cleaning. The DNA concentration was measured with the NanoDrop 8000 V 2.1.0 spectrophotometer and the quality was evaluated on a 1% agarose gel.

Table 1 List of 87 accessions of the *Coffea* genus genotyped by DArTseq

ID in structure	Accession number	Accession	Specie	Source	Origin
1	1	Típica 947	*C. arabica*	Typica selection	Mexico
2	2	Típica Xanthocarpa	*C. arabica*	Typica mutation	Brazil
3	4	Borbón Salvadoreño	*C. arabica*	Bourbon selection	Salvador
4	5	Java	*C. arabica*	Local ethiopian	Indonesia
5	6	Pluma Hidalgo	*C. arabica*	Typica selection	Mexico
6	7	Mundo Novo 23	*C. arabica*	Sumatra × Bourbon	Mexico
7	8	Maragogipe	*C. arabica*	Typica mutation	Brazil
8	9	Pluma Hidalgo 177	*C. arabica*	Typica selection	Mexico
9	10	Blue Mountain	*C. arabica*	Typica selection	Jamaica
10	11	Oro Azteca grano grande	*C. arabica*	Catimor	Mexico
11	12	Limaní	*C. arabica*	Sarchimor	Puerto Rico
12	13	Costa Rica 95	*C. arabica*	Catimor	Costa Rica
13	14	Colombia brote café	*C. arabica*	Catimor	Colombia
14	15	Colombia brote verde	*C. arabica*	Catimor	Colombia
15	16	Catimores T 5175-59 (Masal)	*C. arabica*	Catimor	Portugal
16	17	Colombia Original	*C. arabica*	Catimor	Colombia
17	18	Colombia Amarillo	*C. arabica*	Catimor	Colombia
18	20	IAPAR 59-1	*C. arabica*	Sarchimor	Brazil
19	21	IAPAR Comercial	*C. arabica*	Sarchimor	Paraná
20	22	Catucaí Rojo	*C. arabica*	Icatu × Catucai	Brazil
21	23	Catimores Valdez	*C. arabica*	Catimor	Portugal
22	24	Maracatú 27	*C. arabica*	Marago × Caturra	Brazil
23	25	Oro Azteca	*C. arabica*	Catimor	Mexico
24	26	Amphilo	*C. arabica*	Wild	Ethiopia
25	27	Batie	*C. arabica*	Wild	Ethiopia
26	28	Dessie	*C. arabica*	Wild	Ethiopia
27	29	Filipino	*C. arabica*	Wild	Filipinas
28	30	Geisha	*C. arabica*	Local ethiopian	Ethiopia
29	31	Laurina	*C. arabica*	Bourbon mutation	Java
30	32	Mibirizi	*C. arabica*	Derived from Typica	Congo
31	33	Mokka de Tahití	*C. arabica*	Derived from Mokka (Yemen)	Tahití
32	34	Murta	*C. arabica*	Catimor	Costa Rica
33	35	Purpurascens	*C. arabica*	Typica mutation	Mexico
34	38	Australia Brisbane	*C. arabica*	Wild	Australia
35	39	Chichón PMA	*C. arabica*	Local selection	Mexico
36	40	Cera miel	*C. arabica*	Local selection	Mexico
37	41	Australia Sidney	*C. arabica*	Wild	Australia
38	42	Marsellesa	*C. arabica*	HdT 832/2 × Villa Sarchi CIFC 971/10	Nicaragua
39	43	Castillo Venecia	*C. arabica*	Catimor (multiline)	Colombia
40	45	SL 28	*C. arabica*	Tanganika selection (similar to Bourbon)	Kenya
41	46	Surinam	*C. arabica*	Typica progeny	Surinam
42	47	Icatu 740055-76-8	*C. arabica*	*C. canephora* × Bourbon Vermelho × (BC) Mundo Novo	Brazil
43	48	Variegata sec. E	*C. arabica*	Mutation from *C. arabica*	Mexico

Table 1 continued

ID in structure	Accession number	Accession	Specie	Source	Origin
44	49	Icatu Amarillo	*C. arabica*	*C. canephora* × Bourbon Vermelho × (BC) Mundo Novo	Brazil
45	50	Arabusta	*C. arabica*	Interspecific hybrid	Cameroon
46	52	Caturra Rojo	*C. arabica*	Bourbon mutation	Brazil
47	53	Caturra Amarillo	*C. arabica*	Bourbon mutation	Brazil
48	54	Catuaí Rojo	*C. arabica*	Mundo Novo × Caturra	Brazil
49	55	Catuaí Amarillo	*C. arabica*	Mundo Novo × Caturra	Brazil
50	56	San Roman Amarillo	*C. arabica*	Local selection	Mexico
51	57	Garnica F5	*C. arabica*	Mundo Novo 15 × Caturra Amarillo 13	Mexico
52	58	Garena	*C. arabica*	Garnica mutation	Mexico
53	59	Angustifolia	*C. arabica*	Typica mutation	Mexico
54	60	Castilla	*C. arabica*	Local selection	Mexico
55	61	Pacamara	*C. arabica*	Pacas × Maragogipe	Salvador
56	62	Pacas	*C. arabica*	Bourbon mutation similar to Caturra	Salvador
57	63	Rubí	*C. arabica*	Catuaí × Mundo Novo	Brazil
58	64	San Ramón	*C. arabica*	Mutant	Costa Rica
59	65	Topazio Amarillo	*C. arabica*	Catuaí Amarillo × Mundo Novo	Brazil
60	66	Villalobos	*C. arabica*	Typica mutation	Costa Rica
61	67	Villa Sarchí	*C. arabica*	Bourbon mutation	Costa Rica
62	68	Variegata sec. F	*C. arabica*	Mutation from *C. arabica*	Mexico
63	69	Garnica Iquimite	*C. arabica*	Garnica selection	Mexico
64	70	Caturra Amarillo 13	*C. arabica*	Caturra Rojo mutation	Mexico
65	71	Catuaí Erecta	*C. arabica*	Mundo Novo × Caturra	Brazil
66	73	S 12 Kaffa (Garnica)	*C. arabica*	Mundo Novo × Caturra Amarillo	Ethiopia
67	74	Hibrido de Timor	*C. arabica*	Hybrid	Indonesia
68	75	Acahua	*C. arabica*	Catimor	Brazil
69	76	Catigua	*C. arabica*	Catimor	Brazil
70	77	Paraíso	*C. arabica*	Catimor	Brazil
71	78	Topazio	*C. arabica*	Catuaí × Mundo Novo	Brazil
72	79	Obata	*C. arabica*	Sarchimor × (BC) Catuai	Brazil
73	80	Catucaí Amarillo	*C. arabica*	Icatu × Catuai	Brazil
74	81	H14	*C. arabica*	F1 hybrid by CIRAD	France
75	82	H15	*C. arabica*	*Ídem*	France
76	83	H16	*C. arabica*	*Ídem*	France
77	84	H17	*C. arabica*	*Ídem*	France
78	85	H18	*C. arabica*	*Ídem*	France
79	86	H19	*C. arabica*	*Ídem*	France
80	88	Borbón negro Chiapas	*C. arabica*	Bourbon selection	Mexico
81	89	Geisha Tlacotepec	*C. arabica*	Local ethiopian	Mexico
82	90	Geisha Maduración temprana	*C. arabica*	Local ethiopian	Ethiopia
83	91	Típica naranja	*C. arabica*	Typica selection	Mexico
84	93	Liberica	*C. liberica*	*C. liberica*	Ethiopia
85	94	Robusta Nemaya nuevo	*C. canephora*	*C. canephora* T3561 × *C. canephora* T3751	Brazil

Table 1 continued

ID in structure	Accession number	Accession	Specie	Source	Origin
86	95	Robusta tradicional	*C. canephora*	Local	Mexico
87	96	Robusta Nemaya alto	*C. canephora*	*C. canephora* T3561 × *C. canephora* T3751	Brazil

Catimor Timor hybrid × Caturra, *BC* backcross, *HdT* Timor hybrid

DArTseq analysis based on SNP

For genotypic characterization, the next-generation sequencing technology DArTseq^{TM} was used. DArTseq^{TM} represents a combination of DArT complexity reduction methods, based on methyl filtration and next generation sequencing platforms (Kilian et al. 2012). A genomic representation of the samples was generated by digestion/ligation of the genomic DNA by combination of two restriction enzymes (Pst1-CTGCAG-, HpaII-C/CGG y GGC/C-) and adapters linked by bar code to identify each sample to operate within a single lane on the Illumina HiSeq2500 instrument (Illumina Inc., San Diego, CA).

The site for HpaII was effectively amplified in 30 iterations of PCR, the following reaction profile was used: (1) denaturation at 94 °C for 1 min; (2) 30 cycles [94 °C for 20 s (denaturation), 58 °C for 30 s (pairing by primer) and 72 °C for 45 s (primer's extension)] and (3) final polymerization at 72 °C for 7 min. Equimolar quantities of the amplified fragments were polled by PCR from each reaction of the samples in 96-well microliter plates and applied to the c-Bot bridging PCR (Illumina), followed by sequencing in the Illumina HiSeq2500 system (Illumina Inc., San Diego, CA).

The amplified fragments were successfully sequenced up to 77 base pairs, generating approximately 500,000 unique readings per sample. The analytical program developed and patented by DArT Pvt. Ltd., Australia, was used to generate two types of data, (1) scores for "presence/absence" markers (dominant), called SilicoDArTs (PAVs) and (2) SNP markers. The FASTQ files (full 77 bp readings) were filtered by quality parameters to select high quality markers for this specific study. The PAV markers generated by DArTseq^{TM} were not used in this study.

Data analysis

The dartR package of the R software automatically calculates several quality parameters for each SNP marker, such as call rate, polymorphic information content (PIC) and reproducibility. For the data analysis, the final number of polymorphic SNP markers was taken. The average rate of missing values per

Table 2 Compact grouping of the 87 accessions of *Coffea* spp. product of the dendrogram

Groups	Accessions	Designation
1	33, 28, 59, 26, 73, 49, 55, 24, 76, 17, 30, 1, 9, 66, 2, 64, 8, 27, 5, 45, 89, 29, 35, 38, 60, 70, 74, 7, 69, 41, 50, 4, 67, 11, 32, 78, 65, 80, 46, 52, 53, 61, 71, 58, 63, 31, 62, 48, 57, 34, 40 and 91	Specie: *C. arabica*
2	12, 20, 21, 39, 47, 90, 6, 75, 15, 25, 93, 22, 54, 42, 79, 18 and 77	Species: *C. arabica* and *C. liberica* (93)
3	83, 10, 16, 23, 56, 43, 88, 95, 68, 13 and 14	Species: *C. arabica* and *C. canephora* (95)
4	81, 82, 86, 84 and 85	F1 Hybrid (CIRAD, Francia)
5	94 and 96	Specie: *C. canephora*

marker was 14.7%. Markers with more than 10% of missing values were eliminated, the remaining markers were imputed using the allelic frequencies observed by the statistical software R (R Core Team 2018). The missing genotypes were imputed generating random samples of the marginal distribution of the observed genotypes, that is:

$$x_{ij} \sim Bernoulli(\widehat{p}_j),$$

where $Bernoulli(\widehat{p}_j)$ denotes a Bernoulli random variable with parameter $\widehat{p}_j$ and represents the allelic frequency calculated using the non-missing genotypes (Crossa et al. 2010). Once the markers were imputed, the frequencies of the minor allele (MAF) were obtained and all those markers with MAF $<$ 5% were eliminated. To investigate the relationship between genotypes, a hierarchical grouping based on Euclidean distances and Ward.D2 method between groups, as a measure of similarity (Murthag and Legendre 2014), was performed based on all polymorphic SNP markers. For the heat map, the genomic relations matrix $\boldsymbol{G}$ can be easily calculated using the following expression:

$$\boldsymbol{G} = \frac{\boldsymbol{ZZ}'}{p},$$

where $\boldsymbol{Z}$ is the matrix of markers of dimension n = 87 rows (individuals) and p = 1739 columns (markers), which is obtained by centering and standardizing the columns of the matrix of markers (Kaufman and Rousseeuw 2005; López-Cruz et al. 2015). After, a genlight object was generated by using the dartR and adegenet packages of the R software (Gruber et al. 2017). Then the principal coordinate analysis (PCoA) was performed, PCoA explains the genetic distances among the accessions.

The population structure of the germplasm was analyzed using STRUCTURE v.2.3.4 (Pritchard et al. 2000). The number of hypothetical subpopulations (K) was estimated with the STRUCTURE software through the application of a Bayesian clustering approach for the organization of genetically similar accessions into the same subgroups. A series of Markov Chain Monte Carlo (MCMC) simulations were conducted for each K-value from 1 to 5 with a burn-in length of 10,000, followed by 10,000 iterations. The best K-value was estimated based on the membership coefficient (Q) for each individual in each cluster. The Q values indicate the level of relatedness of each accession to various subgroups.

Results

Genetic-statistical analyses

As a result, 16,995 SNP markers, derived from 34,000 unique sequences, were obtained by DArTseq$^{\mathrm{TM}}$ from 87 accessions of different *Coffea* spp. The number of missing values for the 16,995 SNP markers was found in more than 8000 markers. Most of the markers showed reproducibility $>$ 95%, a call rate $>$ 85% and an average PIC of 0.10. The average of monomorphic markers and missing data were 40.95% and 14.7%, respectively. Because it is not possible to work with missing value rates per marker, an imputation was made based on the frequencies of the observed markers. After removing the markers with more than 10% of the missing data and MAF $<$ 5% (Fig. 1), there were 1739 polymorphic SNP markers for the analysis. The technical and biological replicas allowed evaluating the reliability of the DArTseq$^{\mathrm{TM}}$ method in coffee species.

Clustering analysis

After imputation and elimination of markers based on MAF, a heat map of the 87 accessions was obtained by using the genomic relations matrix $\boldsymbol{G}$ (Fig. 2). Figure 3a and b shows a graph of the first two principal components based on the Euclidean distance matrix, which are identified with different colors in the graph. For PCoA there were 1639 polymorphic SNP markers in the genlight object. PCoA illustrated the genetic divergence among the cultivars and the two main components explain only 32.2% of the total variability. The population distribution determined by these markers is consistent with the output of hierarchical grouping and population structure analysis. *C. arabica* accessions were located in the top two quadrants, while *Coffea liberica* Bull ex. Hiern and *C. canephora* were mainly located in the bottom quadrants.

The accessions of *Coffea* spp. were grouped by the hierarchical method using the Ward.D2 criterion (Murtagh and Legendre 2014) as a measure of proximity between groups (agglomeration method), the resulting dendrogram is shown in Fig. 4. Five well-

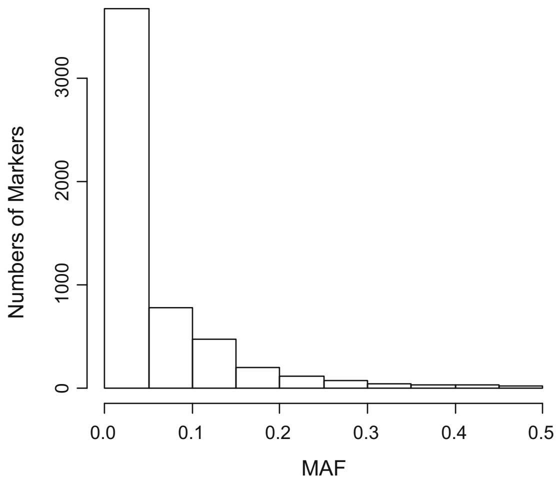

Fig. 1 Frequency distribution of the minor allele (MAF)

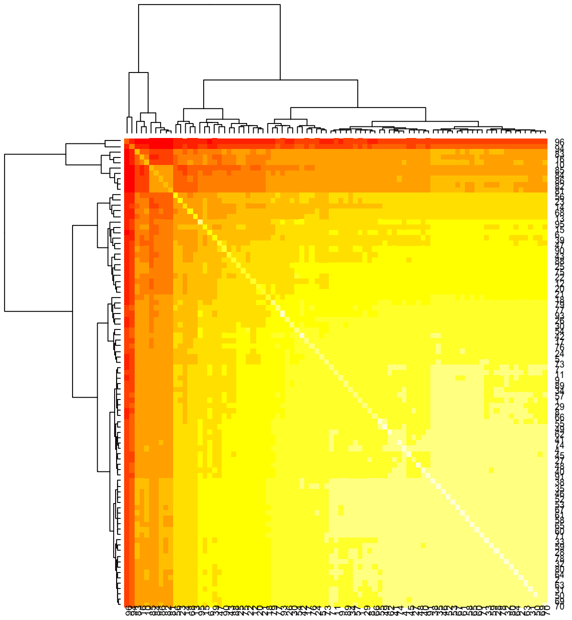

Fig. 2 Heat map for the 87 accessions of *Coffea* spp. from the National Bank of Coffee Germplasm in Mexico using DArTseq Technology

defined groups can be identified in the dendrogram by drawing a horizontal line to cut the tree at a height of 95 (Table 2). The accessions belonging to each group were obtained using the routine "cutree" of the statistical package R (R Core Team 2018). Based on the genomic relations matrix ***G*** and the dendrogram, it was shown that there is genetic diversity among the accessions of *Coffea* spp. and these materials could be promising for use in future breeding programs.

Population structure analysis

The model-based Bayesian cluster analysis in STRUCTURE visualized the population structure under examination (Fig. 5). Five distinct sub-populations were found across cultivars. The sub-populations were denoted as Pop1, Pop2, Pop3, Pop4 and Pop5. The genetic diversity within each sub-population was explained through the estimation of the expected heterozygosity, which varied from 0.07 (Pop2) to 0.28 (Pop1). The expected heterozygosity of Pop3 was 0.09, Pop4 was 0.16 and Pop5 was 0.24. The genetic divergence among the populations revealed by Nei's net nucleotide distance (D) indicated that Pop2 was widely related to the rest of sub-populations, Pop1 (D = 0.34), Pop3 (D = 0.32), Pop4 (D = 0.31) and Pop5 (D = 0.23), respectively. The genetic distance observed between Pop2 and Pop5 (D = 0.18) was the least among the pairs of populations examined (Table 3).

The proportion of membership of individual accessions to each sub-population is illustrated in the bar plot of the population assignment test in structure analysis (Fig. 5). The estimated proportion of membership (Q) suggested that two different species (*C. liberica* [83] and *C. canephora* [85, 86 and 87], red color) were assigned entirely in Pop1. Mostly of *C. arabica* accessions comprised Pop2 (green color). CIRAD F1 hybrids were included in Pop3 (blue color). The remaining accessions showed intermediate and/or highly mixed genetic composition and were hence determined as heterogeneous (Pop4 [yellow color] and Pop5 [pink color]). One CIRAD F1 hybrid accession (76) also shared large amounts of genetic information with Pop4 and Pop5 (Table 3).

Discussion

Genetic-statistical analyses

A total of 1739 SNP markers were used in the present study to provide detailed molecular characterization of 87 accessions of *Coffea* spp. that are in the National Bank of Coffee Germplasm in Mexico. The relationship between genotypes that resulted from different statistical approaches yielded similar results.

The quality parameters of SNP markers in *Coffea* spp. were comparable with other species: watermelon

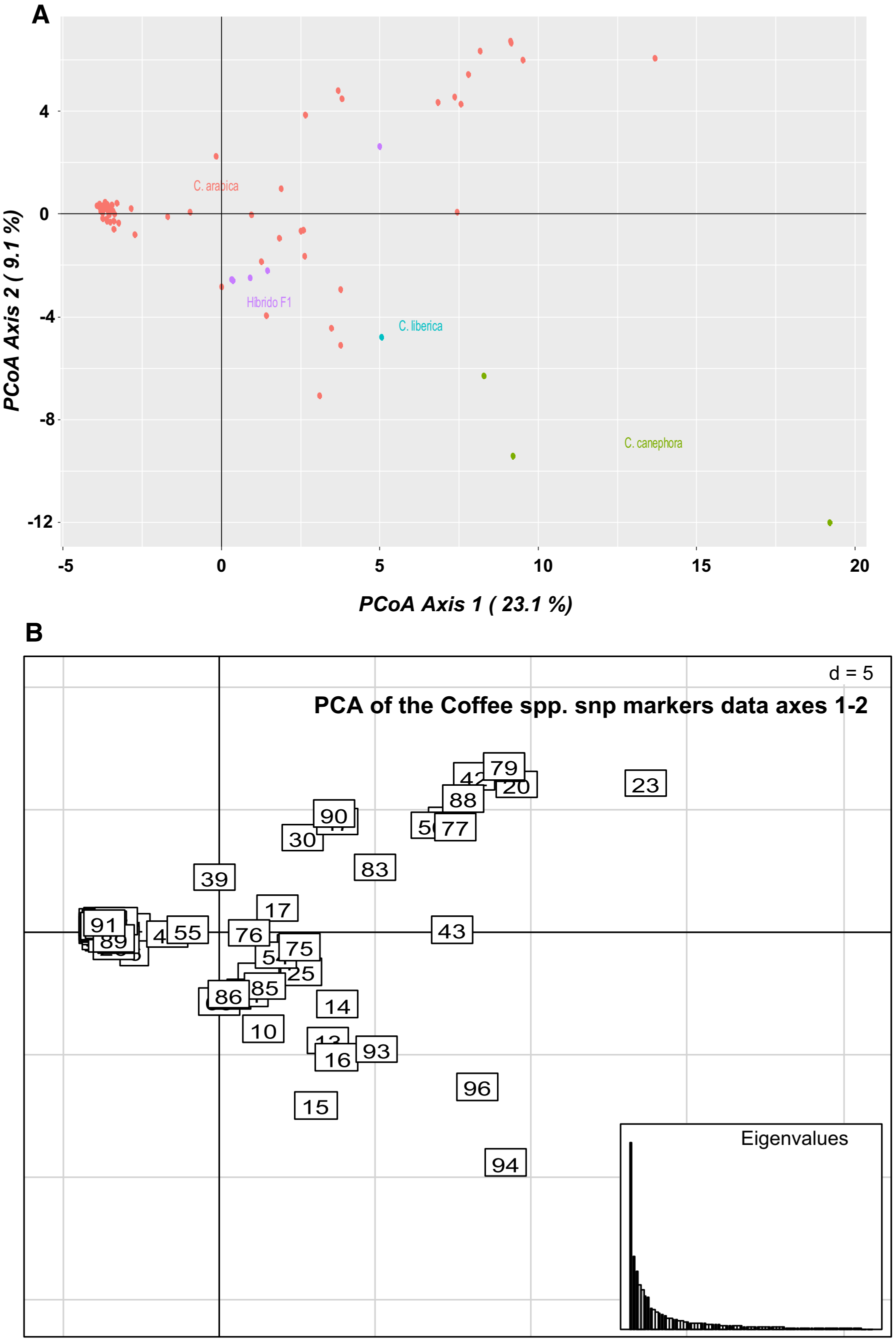

Fig. 3 **a** Principal component 1 vs Principal component 2 of the PCoA explain 32.2% of the variability, **b** PCA represents the grouping for the 87 accessions of *Coffea* spp.

(Yang et al. 2016), *Physaria* spp. (Von Mark et al. 2013), *Sorghum bicolor* (Mace et al. 2008), cassava (Xia et al. 2005) and wheat (Akbari et al. 2006). Based on the polymorphism value, PIC is classified into three categories, high (PIC value higher than 0.5), medium (value between 0.25 and 0.5) and low (lower than 0.25) (Vaiman et al. 1994; Xie et al. 2010). The mean PIC value of the 1739 SNP markers in this population was of 0.10. Moncada and McCouch (2004) also observed low PIC value (0.30) in arabica cultivars using SSR markers. Mishra et al. (2012) obtained the mean PIC values (0.346) in Indian commercial coffee cultivars using polymorphic SRAP markers. Sousa et al. (2017) found a mean PIC value of 0.35 with

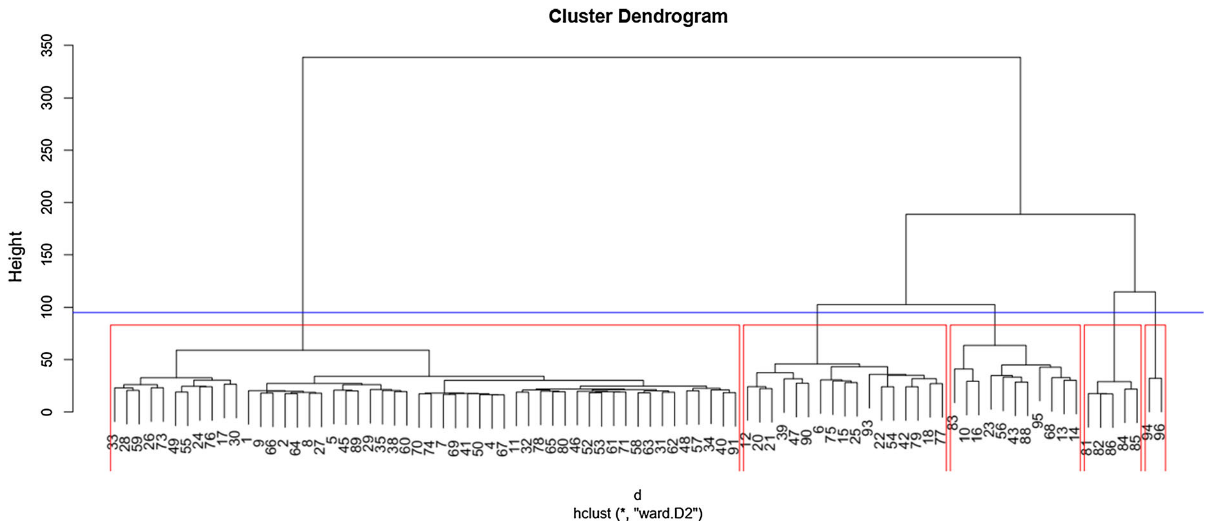

Fig. 4 Dendrogram of 87 accessions of *Coffea* spp. obtained with Euclidean distances calculated from SNP and Ward.D2 method with proximity criterion between groups

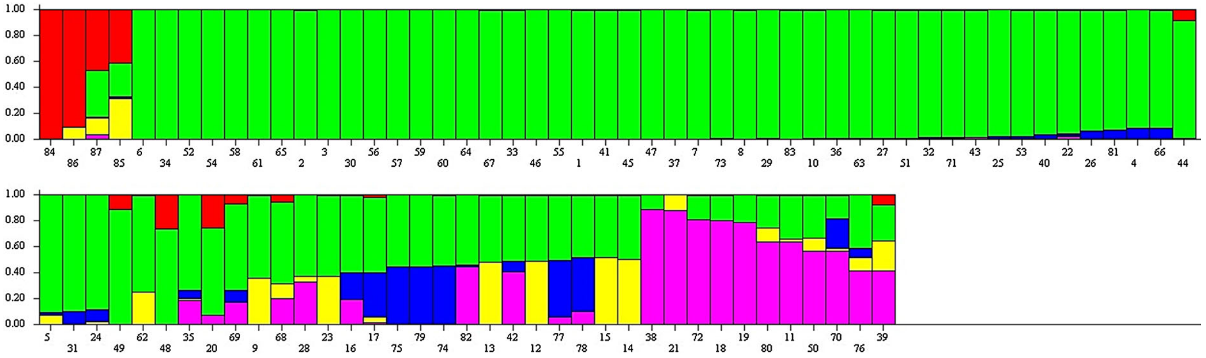

Fig. 5 Population structure of 87 coffee accessions using SNP marker data

Table 3 Genetic divergence among (net nucleotide distance) and within (expected heterozygosity) populations, and the proportion of membership of the population samples

Population	Net nucleotide distance					Expected heterozygosity	Proportion of membership
	Pop1	Pop2	Pop3	Pop4	Pop5		
Pop1	–	0.3444	0.3201	0.3129	0.2326	0.2885	0.043
Pop2		–	0.2597	0.2740	0.1822	0.0745	0.745
Pop3			–	0.3922	0.2998	0.0973	0.046
Pop4				–	0.2711	0.1613	0.053
Pop5					–	0.2466	0.113

11,187 SNP markers. The low PIC value evidences the narrow genetic base of *C. arabica*. The average PIC values of *Coffea* spp. were similar to values identified in SNP markers of watermelon (0.13) and *Physaria* spp. (0.12), but lower than *Sorghum bicolor* (0.41), cassava (0.42), and wheat (0.44).

The SNP markers used in this study have greater abundance and a co-dominant inheritance pattern, which increases their effectiveness in the discrimination of accessions compared to the AFLP, RAPD, SSR and ISSR markers used in previous studies of genetic diversity of coffee. (Lashermes et al. 2011; Garavito et al. 2016; Sant'Ana et al. 2018). Sant'Ana et al. (2018) identified 6696 SNPs from a collection of 107 wild accessions of *C. arabica* from Ethiopia and confirmed great allelic richness in wild accessions, especially in accessions from forests located on the west side of the Great Rift Valley. Sousa et al. (2017) selected 11,187 SNP markers from the coffee population resulting from crosses between the Catuaí and Hybrid of Timor genotypes, the genotyping data provided detailed information on parental genotypes and led to the identification of new candidates as parents for a breeding program.

Our work was done only with a subset of the complete collection of the National Bank of Coffee Germplasm in Mexico. Future studies using the entire collection would be of great value in increasing knowledge about the phenotypic and genotypic diversity of *C. arabica* and related species in Mexico. This study shows that there are genetic differences between *C. arabica* groups, so the selection of genetically diverse parents' lines and exploitation of the heterosis resulting from targeted crosses are promising alternatives in a coffee breeding program.

Clustering analysis

The genomic relations matrix clustering and principal coordinate analysis were used to identify both between and within-species diversity. These analyses grouped the 87 genotypes into five diverse clusters on a principal component plot. The first two components accounted for the 32.2% of the total variation, these data may be understood as follows: there is high genetic distance between *C. canephora* (accessions 94, 95 and 96) and *C. liberica* (93), revealing inter-species diversity. This was shown by Steiger et al. (2002) using AFLP markers. They reported that *C. canephora* and *C. liberica* were more genetically distinct. Finally, it seems to be low genetic distance within *C. arabica* accessions, but in the F1 CIRAD'S hybrids sub population, accession number 83 it's more distant than the rest, could be that belongs to different progenitors. Anagbogu et al. (2019) applied multidimensional scaling (MDS) and found a 36.2% of variation in the re-classification of 46 genotypes of *C. canephora* through genotyping-by-sequencing-single nucleotide polymorphism (GBS-SNP) analysis. Also, the genomic relations matrix ***G*** can be used for studies of the structure of the population of interest or in genomic prediction.

The dendrogram obtained by the Ward.D2's method showed that the 87 genotypes were separated into five dissimilar groups: the first group comprised mostly *C. arabica* genotypes, the second group comprised *C. arabica* genotypes with a *C. liberica* genotype included, the third group comprised a small set of *C. arabica* genotypes and a *C. canephora* genotype was included, the forth group compiles the F1 Hybrids (CIRAD, France) and the fifth comprised two *C. canephora* genotypes. The formation of five distinct groups based on these results made possible to observe that the clustered genotypes form homogeneous groups with similar characteristics and the distinct groups are those among which we find genetic diversity. Bikila et al. (2017) showed genetic diversity in a core collection of 50 *C. canephora* clones and obtained six different groups, which were genotyped with 46,074 SNPs molecular markers.

Population structure analysis

Similar to the dendrogram analysis with previous genotypic characterization of this central collection using SNP markers, population structure analysis, using K = 5, formed five different groups. The first group clustered *C. liberica* and *C. canephora* species, the second group clustered mostly *C. arabica* accessions of the central collection, which evidenced the greater dissimilarity of these accessions with *C. liberica* and *C. canephora* species; the third group clustered CIRAD's F1 hybrids. Also, it was shown by Steiger et al. (2002), using AFLP markers, that *C. canephora* and *C. arabica* were more genetically similar, revealing inter-species diversity even though *C. arabica* resulted from a recent hybridization between *C. canephora* and *C. eugenioides* (Lashermes et al. 1999). Fourth and fifth clusters compiled different *C. arabica* accessions among them. SNP markers and this type of genetic-statistical analysis provide more accurate and less subjective genetic information than that generated from phenotypic data, which is useful in breeding programs (Sousa et al. 2017).

The results obtained from this *Coffea* spp. central collection are similar to those reported in the study of Sant'Ana et al. (2018), who found in the population structure analyses the presence of two to three groups (K = 2 and K = 3), corresponding to the east and west sides of the Great Rift Valley and an additional group formed by wild *C. arabica* accessions collected in the western forests Sousa et al. (2017). analyzed the population structure of coffee genotypes of interest for breeding studies, they used 11,187 SNP markers from which two groups (K = 2) were obtained.

Conclusion

DArTseq^{TM} technology identified 1739 SNP polymorphic markers, which discriminated five divergent groups at a distance of 95 and detected low genetic variation among the *Coffea* spp. of the central collection. The identified groups have promising genotypes within them and could be useful for the establishment of a coffee breeding program in Mexico. Our study confirmed that the genotyping method by DArTseq^{TM} can be successfully used in studies of genetic diversity.

Funding This work was supported by the FONDO SECTORIAL SAGARPA-CONACYT [2016-01-277838].

Compliance with ethical standards

Conflict of interest The authors have no conflicting interests, and all authors have approved the manuscript and agree with its submission to the Genetic Resources and Crop Evolution—An International Journal.

Ethical approval This material has not been published in whole or in part elsewhere; the manuscript is not currently being considered for publication in another journal; all authors have been personally and actively involved in substantive work leading to the manuscript, and will hold themselves jointly and individually responsible for its content.

References

Akbari M, Wenzl P, Caig V, Carling J, Xia L, Yang S et al (2006) Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. Theor Appl Genet 113(8):1409–1420. https://doi.org/10.1007/s00122-006-0365-4

Anagbogu CF, Bhattacharjee R, Ilori C et al (2019) Genetic diversity and re-classification of coffee (*Coffea canephora* Pierre ex A. Froehner) from South Western Nigeria through genotyping-by-sequencing-single nucleotide polymorphism analysis. Genet Resour Crop Evol 66:685. https://doi.org/10.1007/s10722-019-00744-2

Avelino J, Cristancho M, Georgiou S, Imbach P, Aguilar L, Bomemann G, Läderach P, Anzueto F, Hruska AJ, Morales C (2015) The coffee rust crises in Colombia and Central America (2008–2013): impacts, plausible causes and proposed solution. Food Secur 7:313–321

Berdugo-Cely J, Valbuena RI, Sánchez BE, Barrero LS, Yockteng R (2017) Genetic diversity and association mapping in the Colombian Central Collection of *Solanum tuberosum* L. Andigenum group using SNPs markers. PLoS ONE 12(3):e0173039. https://doi.org/10.1371/journal.pone.0173039

Bikila BA, Sakiyama NS, Caixeta ET (2017) SNPs based molecular diversity of *Coffea canephora*. J Microbiol Exp 5(1):00136. https://doi.org/10.15406/jmen.2017.05.00136

Carvalho A, Krug CA (1949) Agentes de polinização da flor do cafeeiro (*Coffea arabica* L.). Bragantia 9(4):11–24. https://doi.org/10.1590/S0006-87051949000100002

Crossa J, de los Campos G, Pérez-Rodríguez P, Gianola D, Burgueño J et al (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724

Cruz VMV, Kilian A, Dierig DA (2013) Development of DArT marker platforms and genetic diversity assessment of the US collection of the new oilseed crop lesquerella and related species. PLoS ONE 8:e64062. https://doi.org/10.1371/journal.pone.0064062

Cubry P, Musoli P, Legnate H et al (2008) Diversity in coffee assessed with SSR markers: structure of the genus coffea and perspectives for breeding. Genome 51:50–63. https://doi.org/10.1139/G07-096

Davis AP, Tosh J, Ruch N, Fay M (2011) Growing coffee: *Psilanthus* (Rubiaceae) subsumed on the basis of molecular and morphological data; implications for the size, morphology, distribution and evolutionary history of *Coffea*. Bot J Linn Soc 167:357–377. https://doi.org/10.3389/fpls.2017.00138

Egea LA, Mérida-García R, Kilian A, Hernández P, Dorado G (2017) Assessment of genetic diversity and structure of large garlic (*Allium sativum*) germplasm bank, by diversity arrays technology "Genotyping-by-Sequencing" platform (DArTseq). Front Genet 8:98. https://doi.org/10.3389/fgene.2017.00098

Escamilla E, Ruiz O, Díaz G, Landeros C, Platas DE, Zamarripa A, González VA (2005) El agroecosistema café orgánico en México. Manejo Integr Plagas Agroecol 76:5–16

Ferrão LFV, Caixeta ET, Pena G et al (2015) New EST–SSR markers of *Coffea arabica*: transferability and application to studies of molecular characterization and genetic mapping. Mol Breed 35:31. https://doi.org/10.1007/s11032-015-0247-z

Garavito A, Montagnon C, Guyot R, Bertrand B (2016) Identification by the DArTseq method of the genetic origin of the *Coffea canephora* cultivated in Vietnam and Mexico. BMC Plant Biol 16:242. https://doi.org/10.1186/s12870-016-0933-y

Gruber BU, Berry P, Georges OA (2017) dartR: an R package to facilitate analysis of SNP data generated from reduced representation genome sequencing. Mol Ecol Resour. https://doi.org/10.1111/1755-0998.12745

Hoisington D, Khairallah M, Gonzalez-de-Leon D (1994) Laboratory protocols, CIMMYT applied molecular genetics laboratory, 2nd edn. CIMMYT, Mexico

ICO (2017) International coffee organization. https://www.ico.org/prices/po-production.pdf. Accessed 6 April 2017.

Kaufman L, Rousseeuw PJ (2005) Finding groups in data: an introduction to cluster analysis. Wiley, New York

Kilian A, Wenzl P, Huttner E, Carling J, Xia L, Blois H et al (2012) Diversity arrays technology: a generic genome profiling technology on open platforms. Methods Mol Biol 888:67–89. https://doi.org/10.1007/978-1-61779-870-2_5

Lashermes P, Combes MC, Robert J et al (1999) Molecular characterization and origin of the *Coffea arabica* L. genome. Mol Gene Genet 261:259–266. https://doi.org/10.1007/s004380050965

Lashermes P, Combes MC, Ansaldi C et al (2011) Analysis of alien introgression in coffee tree (*Coffea arabica* L.). Mol Breed 27:223–232. https://doi.org/10.1007/s11032-010-9424-2

López-Cruz M, Crossa J, Bonnet D, Dreisigacker S, Poland J, Jannink LL, Singh RP, Autrique E, de los Campos G (2015) Increased prediction accuracy in wheat breeding trials using a marker × environment interaction genomic selection model. G3. https://doi.org/10.1534/g3.114.016097

López-García FJ, Escamilla-Prado E, Zamarripa-Colmenero A, Cruz-Castillo JG (2016) Producción y calidad en variedades de café (*Coffea arabica* L.) en Veracruz, Mexico. Rev Fitotec Mex 39(3):297–304

Mace ES, Xia L, Jordan DR, Halloran K, Parh DK, Huttner E et al (2008) DArT markers: diversity analyses and mapping in *Sorghum bicolor*. BMC Genom 9(1):1

Mishra MK, Sandhyarani N, Suresh N, Satheesh Kumar S, Soumya PR, Yashodha MH, Bhat A (2012) Genetic diversity among indian coffee cultivars determined via molecular markers. J Crop Improv 26(6):727–750. https://doi.org/10.1080/15427528.2012.696085

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

Moncada MDP, Tovar E, Montoya JC et al (2016) A genetic linkage map of coffee (*Coffea arabica* L.) and QTL for yield, plant height, and bean size. Tree Genet Genom 12:5. https://doi.org/10.1007/s11295-015-0927-1

Moreno RLG (2004) Obtención de variedades de café con resistencia durable a enfermedades, usando la diversidad genética como estrategia de mejoramiento. Rev Acad Colomb Cienc Exact Físic Nat 28(107):187–200

Murtagh F, Legendre P (2014) Ward's hierarchical agglomerative clusterion method: which algorithms implement Ward's criterion? J Classif 31:274–295. https://doi.org/10.1007/s00357-014-9161-z

Pailles Y, Ho S, Pires IS, Tester M, Negrão S, Schmöckel SM (2017) Genetic diversity and population structure of two tomato species from the Galapagos Islands. Front Plant Sci 8:138

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

R Core Team (2018) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Raman H, Raman R, Kilian A, Detering F, Carling J, Coombes N et al (2014) Genome-wide delineation of natural variation for pod shatter resistance in *Brassica napus*. PLoS ONE 9:e101673. https://doi.org/10.1371/journal.pone.0101673

Sant'Ana GC, Pereira LFP, Pot D, Ivamoto ST, Domingues SD, Ferreira RV, Leroy T (2018) Genome-wide association study reveals candidate genes influencing lipids and diterpenes contents in *Coffea arabica* L. Sci Rep 8:465. https://doi.org/10.1038/s41598-017-18800-1

Sousa TV, Caixeta ET, Alkimim ER, Baião de Oliveira AC, Pereira AA, Sakiyama NS, de Resende JMFR, Zambolim L (2017) Population structure and genetic diversity of coffee progenies derived from Catuaí and Híbrido de Timor revealed by genome-wide SNP marker. Tree Genet Genome 13:124. https://doi.org/10.1007/s11295-017-1208-y

Steiger DL, Nagai C, Moore PH, Morden CW, Osgood RV, Ming R (2002) AFLP analysis of genetic diversity within and among *Coffea arabica* cultivars. Theor Appl Genet 105(2–3):209–215. https://doi.org/10.1007/s00122-002-0939-8

Vaiman D, Mercier D, Moazai G (1994) A set of 99 cattle microsatellite, characterization, synteny mapping and polymorphism. Mamm Genome 5:288–297

Van Hintum ThJL, Brown AHD, Spillane C, Hodgkin (2000) Core collections of plant genetic resources. IPGRI Tech Bull 3

Von Mark VC, Kilian A, Dierig DA (2013) Development of DArT marker platforms and genetic diversity assessment of the US collection of the new oilseed crop lesquerella and related species. PLoS ONE 8(5):e64062. https://doi.org/10.1371/journal.pone.0064062

Xia L, Peng K, Yang S, Wenzl P, De Vicente MC, Fregene M et al (2005) DArT for high-throughput genotyping of cassava (*Manihot esculenta*) and its wild relatives. Theor Appl Genet 110(6):1092–1098. https://doi.org/10.1007/s00122-005-1937-4

Xie WG, Zhang XQ, Cai H, Liu W, Peng Y (2010) Genetic diversity analysis and transferability of cereal EST-SSR markers to orchard grass (*Dactylis Glomerata* L.). Biochem Syst Ecol 38(4):740–749. https://doi.org/10.1016/j.bse.2010.06.009

Yang X, Ren R, Ray R, Xu J, Li P, Zhang M et al (2016) Genetic diversity and population structure of core watermelon (*Citrullus lanatus*) genotypes using DArTseq-based SNPs. Plant Genet Resour 14(03):226–233

Zhou L, Vega FE, Tan H, Lluch AER, Meinhardt LWF, Mischke WS, Irish B, Zhang D (2016) Developing single nucleotide polymorphism (SNP) markers for the identification of coffee germplasm. Trop Plant Biol 9:82–95. https://doi.org/10.1007/s12042-016-9167-2

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.