Scientific Reports

https://doi.org/10.1038/s41598-026-35855-1

Article in Press

Genomic discrimination of the botanical groups conilon and robusta of *Coffea canephora*

Received: 21 October 2025

Accepted: 8 January 2026

Published online: 16 January 2026

Cite this article as: Oliveira R.G., Almeida F.A.N., Zaidan I.R. *et al.* Genomic discrimination of the botanical groups conilon and robusta of *Coffea canephora*. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-35855-1

Rosana Gomes Oliveira, Francine Alves Nogueira Almeida, Iasmine Ramos Zaidan, João Felipe Brites Senra, Matheus Alves Silva, Adésio Ferreira & Marcia Flores Silva Ferreira

GENOMIC DISCRIMINATION OF THE BOTANICAL GROUPS CONILON AND ROBUSTA OF *Coffea canephora*

Rosana Gomes de Oliveira[1], Francine Alves Nogueira de Almeida[1], Iasmine Ramos Zaidan[1], João Felipe de Brites Senra[2], Matheus Alves Silva[1], Adésio Ferreira[1], Marcia Flores da Silva Ferreira[1]*

[1]Laboratory of Genetics and Plant Breeding, Department of Agronomy, Center for Agricultural Sciences and Engineering, Federal University of Espírito Santo, Alegre, ES, Brazil. [2]Capixaba Institute for Research, Technical Assistance and Rural Extension, Cachoeiro de Itapemirim, ES, Brazil.

*Corresponding author
E-mail address: marcia.ferreira@ufes.br
Federal University of Espírito Santo, Department of Agronomy, Alegre, Espírito Santo. Brazil, Zip Code: 29500-000.

Orcid dos authors:
Rosana Gomes de Oliveira: https://orcid.org/0009-0000-7325-0666
Francine Aves Nogueira de Almeida: https://orcid.org/0000-0001-7996-6227
Iasmine Ramos Zaida: https://orcid.org/0000-0001-9115-4504
João Felipe de Brites Senra: https://orcid.org/0000-0001-7915-2821
Matheus Alves da Silva: https://orcid.org/0000-0002-4992-9588
Adésio Ferreira: https://orcid.org/0000-0002-7000-1725
*Marcia Flores da Silva Ferreira: https://orcid.org/0000-0003-1541-6634

Abstract

The correct identification of *Coffea canephora* genotypes into the botanical groups Conilon, Robusta and their hybrids is fundamental for breeding strategies aimed at developing and selecting commercial clones; however, genotype discrimination is challenging when based only on morphological attributes. A total of 121 individuals were evaluated from the Incaper Active Germplasm Bank (BAG): 52 Conilon, 33 Robusta, and 36 hybrids (morphological intermediates showing traits of both groups). DArTseq™ SNPs were filtered, and 1,551 markers were used in the analyses. Cluster analyses revealed three genetic groups: one Robusta (G1) and two Conilon (G2 and G3). AMOVA indicated that 78.49% of the variation occurred among groups and 21.51% within groups. Differentiation was greatest between G1 and G3 (F_{st} = 0.77), followed by G1 and G2 (F_{st} = 0.46), whereas G2 and G3 showed lower differentiation (F_{st} = 0.31). G2 showed high genetic diversity and greater genotypic variability, with three SNP-defined genotypes, suggesting the presence of genetic hybrids. Ten SNPs, located on nine chromosomes and outside genic regions, were validated to discriminate the botanical groups, providing a molecular fingerprinting tool for the characterization of *C. canephora* germplasm in Brazil.

Keywords: *Coffea* spp., Genetic diversity, Population structure, Molecular markers.

Introduction

Conilon and Robusta are the world's main cultivated botanical groups of *C. canephora* [1]. They have phenotypic and genetic differences that allow the exploration of heterosis in hybrid populations [1, 2]. Conilon plants are reported to have bushy growth, elongated leaves, early maturation, and greater tolerance to drought [3, 4, 5]. Meanwhile, Robusta plants present erect growth, larger leaves, late maturation, greater resistance to pests and diseases, and lower tolerance to water stress [3, 5].

The discrimination and characterization of Conilon and Robusta plants have been carried out through genealogical studies and morphological and genetic descriptors [1, 2, 6, 7, 8]. However, differentiation can be difficult due to morphological traits, since individuals with phenotypic characteristics of both groups are often designated as hybrids [9, 10, 11]. In this regard, the research unit of the Brazilian Agricultural Research Corporation (Embrapa) in the state of Rondônia has stood out in the development of commercial hybrids between Conilon and Robusta, with the aim of combining agronomic characteristics of interest in both groups [12].

With respect to the genetic groups of *C. canephora*, two germplasm groups were initially described using molecular markers: Congolese and Guinean. The Congolese group has two subdivisions, SG1 and SG2, in which SG1 refers to Conilon, widely cultivated in the Brazilian state of Espírito Santo, while SG2 refers to Robusta, cultivated worldwide [2, 3, 13, 14, 15, 16]. Based on genotyping involving the Coffee 8.5K SNP array, a new classification into eight well-differentiated genetic groups was proposed, corresponding to different geographic origins (A, B, C, D, E, G, O and R) [17]. In this classification, Conilon is in group A while Robusta is in groups B and E [14, 16, 17]. However, in Brazil, the world's largest coffee producer [18], comparative studies among cultivated botanical groups are scarce, involving intra-and interpopulation improvement strategies [10, 12, 19].

In Espírito Santo, the country's leading grower of *C. canephora* [8], the germplasm bank of the Capixaba Institute for Research, Technical Assistance and Rural Extension (Incaper) has approximately 600 accessions classified as Conilon, Robusta or hybrids (morphological intermediates), based on 29 morphological and phenological descriptors [10, 19], as defined by the Ministry of Agriculture, Animal Husbandry and Supply (MAPA) for the genus *Coffea* spp. and interspecific hybrids. One limitation of this classification is that the reference cultivars for these descriptors are all *Coffea arabica*, so there are no specific references for *C. canephora* genotypes [20], as shown in studies carried out with Incaper accessions [10, 19]. In addition, morphological classification is influenced by environmental variation and by the set of traits used, whereas DNA molecular markers are free from environmental effects, can be applied to seeds or early developmental stages, and may reduce time and costs for discrimination tests, provided that an appropriate regulatory framework is in place [6]. Furthermore, with the increasing number of clones, performing DUS (Distinctness, Uniformity, and Stability) tests is becoming increasingly costly [21].

Among the various studies of the genetic diversity and population structure of *C. canephora* and other *Coffea* species, the use of molecular markers such as single nucleotide polymorphisms (SNPs), simple sequence repeats (SSRs), inter-simple sequence repeats (ISSRs) and amplified fragment length polymorphism (AFLP) has been reported. Among these, SNPs stand out for being abundant in the genome, biallelic and codominant, in addition to allowing automated genotyping by next generation sequencing technologies, resulting in time and cost savings [22, 23]. The DArTseq™ method enables the rapid identification of thousands of high-quality SNPs in predominantly genic regions [24] and has been used in studies of diversity and population structure of *C. canephora*, as well as in genome-wide selection and genome-wide association studies [2, 22, 23, 24, 25].

We hypothesized that the prior classification of Incaper accessions based solely on morphological traits has limitations to consistently represent the genetic groups of *C. canephora*. Therefore, we aimed to characterize

genetic diversity and population structure within and between the botanical groups Conilon and Robusta using SNP markers and, additionally, to identify a minimal set of SNPs with potential to discriminate genetic groups of *C. canephora*, analyzing 121 previously classified individuals as Conilon (n = 52), Robusta (n = 33) and hybrids (n = 36; morphological intermediates).

Materials and Methods

Plant Materials

We selected 121 individuals from the 600 *C. canephora* accessions in the Incaper BAG [10, 19], comprising 52 Conilon, 33 Robusta, and 36 morphological intermediates (hybrids) (Table 1). The accessions of each group were chosen to represent their genetic diversity and agronomic relevance, in collaboration with Incaper researchers during the planning of the present study. Of these 121 individuals, 27 are clones belonging to the commercial cultivars Andina, Centenário ES8132, Diamante ES8112, EMCAPA 8141 – Robustão Capixaba, Jequitibá – ES8122, Marilândia – ES8143, Monte Pascoal, Plena, Tributun and Vitória Incaper 8142. The Incaper BAG accessions contain progenies selected from crops in different municipalities of Espírito Santo and southern Bahia, introductions from Robusta collections of the Agronomic Institute of Campinas (IAC) and the Minas Gerais Agricultural Research Corporation (EPAMIG), and progenies from controlled crosses carried out by researchers of Incaper [10].

The detailed classification of BAG accessions into the botanical groups Conilon, Robusta and hybrids was described in the works of Ferrão *et al*. [10, 19], based on morphological studies of 29 morphoagronomic descriptors of the species, according to MAPA [20]. The averages of each descriptor, by group of genotypes studied, are presented in Table 2. The descriptors used were related to the overall characteristics of the plant, as well as branches, leaves, fruits, and seeds, as well as maturation cycle and reaction to pests, diseases, and drought, which were evaluated categorically. The phenotypic evaluations of the plants were conducted in Espírito Santo at the experimental farms of Marilândia (FEM) and Bananal do Norte (FEBN), from May to December 2019, in plants aged between 24 and 30 months. FEM is in the municipality of Marilândia, in the northwest region of the state (19.407°S latitude and 40.539°W longitude), with average annual temperature of 24.4 °C, average annual precipitation of 1150 mm, and altitude of 95 m. Each accession is represented by a plot of five plants, with spacing of 3.5 x 1.20 m. In turn, FEBN is in the municipality of Cachoeiro de Itapemirim (20.750°S latitude and 41.228°W longitude), with average annual temperature of 24.1 °C, average annual rainfall of 1200 mm, and altitude of 50 m. Each accession is represented by a plot of three plants, with spacing of 3.0 x 1.20 m [10].

Table 1. Identification of the 121 individuals from the Incaper BAG, including classification by genetic group, differentiation by morphological traits and the inclusion of clones of commercial cultivars.

Clone*	Genetic group**	Morphological group***	Commercial cultivars****	MAPA registration*****
R2[1]	G2	Conilon	Robustão Capixaba	5385
R3[1]	G3	Conilon	Robustão Capixaba	5385
R4[1]	G3	Conilon	Robustão Capixaba	5385
R5[1]	G2	Conilon	Robustão Capixaba	5385
R10[1]	G3	Conilon	Robustão Capixaba	5385
V01[1]	G3	Conilon	Vitória	20471
V06[2]	G3	Conilon	Vitória	20471
V07[2]	G2	Conilon	Vitória	20471
V08[2]	G3	Conilon	Vitória	20471
V10[2]	G3	Conilon	Vitória	20471

101[1]	G3	Conilon	Diamante	31002
201/LB1[1]	G3	Conilon	Jequitibá, Monte Pascoal e Plena	31003, 44082, 50300
301[1]	G3	Conilon	Centenário	31001
306[1]	G3	Conilon	Centenário	31001
308[1]	G3	Conilon	Centenário	31001
309[1]	G3	Conilon	Centenário	31001
406[1]	G3	Conilon	Centenário	31001
407[1]	G3	Conilon	Marilândia	31001
01 Sul Bahia[1]	G3	Conilon	-	-
10 Sul Bahia[1]	G3	Conilon	-	-
13 Sul Bahia[1]	G3	Conilon	-	-
30 Sul Bahia[1]	G3	Conilon	-	-
33 Sul Bahia[1]	G3	Conilon	-	-
BAG33[1]	G3	Conilon	-	-
BAG35[1]	G3	Conilon	-	-
BAG38[2]	G3	Conilon	-	-
BAG57[1]	G3	Conilon	-	-
BAG58[1]	G3	Conilon	-	-
BAG59[1]	G3	Conilon	-	-
BAG93[1]	G3	Conilon	-	-
BAG110[1]	G2	Conilon	-	-
BAG119[1]	G3	Conilon	-	-
BAG127[1]	G3	Conilon	-	-
BAG129[1]	G3	Conilon	-	-
BAG138[1]	G2	Conilon	-	-
BAG196[2]	G3	Conilon	-	-
BAG202[1]	G3	Conilon	-	-
BAG216[2]	G3	Conilon	-	-
BAG223[2]	G3	Conilon	-	-
BAG267[1]	G3	Conilon	-	-
BAG273[1]	G3	Conilon	-	-
BAG279[1]	G3	Conilon	-	-
BAG291[1]	G3	Conilon	-	-
BAG319[1]	G3	Conilon	-	-
BAG320[1]	G3	Conilon	-	-
BAG327[1]	G3	Conilon	-	-
BAG328[1]	G2	Conilon	-	-
BAG341[1]	G3	Conilon	-	-
BAG347[1]	G3	Conilon	-	-
BAG349[2]	G3	Conilon	-	-
BAG354[1]	G3	Conilon	-	-
BAG359[1]	G3	Conilon	-	-
EP03[2]	G1	Robusta	-	-
EP07[2]	G1	Robusta	-	-
EP13[2]	G1	Robusta	-	-
EP20[1]	G1	Robusta	-	-
EP23[1]	G1	Robusta	-	-
EP38[2]	G1	Robusta	-	-
EP60[2]	G1	Robusta	-	-
EP63[2]	G1	Robusta	-	-
IAC01[1]	G1	Robusta	-	-
IAC02[2]	G1	Robusta	-	-
IAC05[1]	G1	Robusta	-	-
IAC09[1]	G1	Robusta	-	-

IAC10[1]	G1	Robusta	-	-
IAC11[1]	G1	Robusta	-	-
IAC17[2]	G1	Robusta	-	-
IAC19[2]	G1	Robusta	-	-
IAC21[1]	G1	Robusta	-	-
IAC22[1]	G1	Robusta	-	-
IAC23[1]	G1	Robusta	-	-
IAC27[1]	G1	Robusta	-	-
IAC32[1]	G1	Robusta	-	-
IAC35[2]	G1	Robusta	-	-
IAC38[1]	G1	Robusta	-	-
IAC45[2]	G1	Robusta	-	-
IAC50[2]	G1	Robusta	-	-
IAC51[2]	G2	Robusta	-	-
IAC55[1]	G1	Robusta	-	-
IAC56[1]	G1	Robusta	-	-
IAC71[2]	G1	Robusta	-	-
IAC72[1]	G1	Robusta	-	-
IAC73[1]	G1	Robusta	-	-
IAC74[1]	G1	Robusta	-	-
BAG558[2]	G2	Robusta	-	-
A1[2]	G2	Hybrid	Andina, Plena, Tributum e Diamante	31002, 39441, 50300
203[1]	G2	Hybrid	Jequitiba	31003
207[1]	G2	Hybrid	Jequitibá	31003
208[1]	G3	Hybrid	Jequitibá	31003
405[1]	G3	Hybrid	Marilândia	37678
V03[2]	G2	Hybrid	Vitória	20471
V05[2]	G3	Hybrid	Vitória	20471
V11[2]	G3	Hybrid	Vitória	20471
V13[2]	G3	Hybrid	Vitória	20471
EP22[2]	G1	Hybrid	-	-
EP31[2]	G1	Hybrid	-	-
EP35[2]	G1	Hybrid	-	-
EP57[2]	G1	Hybrid	-	-
09Robustão[2]	G3	Hybrid	-	-
25 Sul Bahia[1]	G3	Hybrid	-	-
IAC18[2]	G1	Hybrid	-	-
IAC39[2]	G1	Hybrid	-	-
IAC48[2]	G1	Hybrid	-	-
BAG14[1]	G3	Hybrid	-	-
BAG42[1]	G3	Hybrid	-	-
BAG61[1]	G3	Hybrid	-	-
BAG88[1]	G3	Hybrid	-	-
BAG98[1]	G3	Hybrid	-	-
BAG111[1]	G3	Hybrid	-	-
BAG134[1]	G3	Hybrid	-	-
BAG205[1]	G3	Hybrid	-	-
BAG238[1]	G3	Hybrid	-	-
BAG242[2]	G3	Hybrid	-	-
BAG254[1]	G3	Hybrid	-	-
BAG255[1]	G3	Hybrid		
BAG312[1]	G3	Hybrid	-	-
BAG367[1]	G3	Hybrid	-	-
BAG380[1]	G3	Hybrid	-	-

BAG384[1]	G3	Hybrid	-	-
BAG361[2]	G3	Hybrid	-	-
BAG385[1]	G3	Hybrid	-	-

*Collection location: [1]FEBN (Bananal do Norte – ES); [2]FEM (Marilândia – ES). **Genetic group based on SNPs: G1 (Robusta), G2 (Hybrids), and G3 (Conilon). ***Botanical group based on morphological traits [10, 19]. ****Commercial cultivars. ****MAPA registration numbers [20].

Table 2. Means and standard deviations (SD) of the traits of each group of plants used: Conilon, Robusta and Hybrids. (N = number of plants).

Traits	Conilon (N = 52)	Robusta (N = 33)	Hybrid (N = 36)
Shape*	1.92 ± 0.25	2.00 ± 0.00	2.00 ± 0.00
Height*	2.83 ± 0.49	4.45 ± 0.56	3.22 ± 0.68
Crown diameter*	3.72 ± 0.55	4.77 ± 0.42	4.16 ± 0.50
Internode length*	1.94 ± 0.35	2.94 ± 0.23	2.22 ± 0.48
Plagiotropic branch intensity*	2.51 ± 0.57	2.51 ± 0.50	2.38 ± 0.49
Leaf length*	2.08 ± 0.35	2.88 ± 0.32	2.50 ± 0.50
Leaf width*	2.03 ± 0.37	2.71 ± 0.45	2.22 ± 0.42
Leaf shape*	2.60 ± 0.80	2.71 ± 0.45	2.27 ± 0.94
Young leaf color*	2.64 ± 0.48	1.68 ± 0.96	2.58 ± 0.55
Adult leaf color*	1.98 ± 0.13	2.42 ± 0.88	2.00 ± 0.00
Edge ondulation*	2.14 ± 0.40	2.00 ± 0.00	2.19 ± 0.52
Secondary veins*	1.87 ± 0.50	2.28 ± 0.51	2.22 ± 0.54
Fruit size*	3.14 ± 0.58	2.42 ± 0.55	3.33 ± 0.67
Fruit shape*	2.19 ± 0.58	3.42 ± 0.55	2.13 ± 0.59
Ripe fruit color*	2.85 ± 0.35	1.17 ± 0.38	3.16 ± 0.50
Fruit adhesion degree*	1.50 ± 0.60	3.54 ± 0.50	1.52 ± 0.60
Mesocarp succulence*	2.03 ± 0.37	1.51 ± 0.74	2.25 ± 0.55
Seed length*	1.83 ± 0.70	2.37 ± 0.54	1.91 ± 2.01
Seed width*	1.82 ± 0.66	2.68 ± 0.52	2.11 ± 1.98
Seed thickness*	1.98 ± 0.55	2.71 ± 0.45	2.13 ± 1.97
Endosperm color*	1.28 ± 0.45	2.00 ± 0.00	1.86 ± 0.35
Tonality of the coverage film*	1.21 ± 0.41	2.00 ± 0.00	1.86 ± 0.35
Degree of pellicle adherence*	1.55 ± 0.60	1.77 ± 0.59	1.50 ± 0.65
Ripening cycle	2.57 ± 0.78	3.08 ± 0.98	2.27 ± 2.03
Rust	3.96 ± 0.60	1.08 ± 0.28	2.61 ± 1.43
Berryblotch	3.07 ± 0.42	2.97 ± 0.38	2.91 ± 0.64
Leaf miner	2.87 ± 0.38	3.80 ± 0.63	3.05 ± 0.58
Cochineal	2.03 ± 0.38	1.77 ± 0.42	2.05 ± 0.89
Drought tolerance	1.37 ± 0.55	2.28 ± 0.45	1.75 ± 0.64
Sc/ha	33.15 ± 11.02	21.43 ± 7.78	33.92 ± 11.36
Floating cherries %	18.26 ± 14.09	17.75 ± 10.69	17.38 ± 12.60
Cherry/coconut	2.31 ± 0.18	2.51 ± 0.19	2.38 ± 0.18
Coconut/processed coffee	1.91 ± 0.21	2.43 ± 0.40	2.06 ± 0.33
Cherry/processed coffee	4.35 ± 0.57	6.06 ± 1.24	4.77 ± 0.86
Screen 17	3.01 ± 5.02	11.93 ± 10.80	5.15 ± 6.51
Screen 15	21.85 ± 14.73	24.50 ± 10.14	25.61 ± 11.78
Screen 13	31.44 ± 12.42	24.65 ± 11.73	26.48 ± 10.57
Screen 12	2.13 ± 3.04	4.71 ± 4.89	3.14 ± 2.38
Screen 11	10.02 ± 6.51	11.03 ± 6.43	13.14 ± 5.42
Screen 10	14.58 ± 6.96	10.51 ± 5.36	13.35 ± 6.88
Greater than 15	24.87 ± 18.25	36.43 ± 17.75	30.77 ± 14.68
Greater than 13	53.31 ± 15.15	61.08 ± 12.59	57.24 ± 11.16
Screens 10–12	26.78 ± 11.12	26.25 ± 11.31	26.63 ± 9.89
Mean	16.92 ± 13.36	12.67 ± 10.44	13.13 ± 8.33

Based on the data provided by Incaper and the results presented in the articles by Ferrão *et al*. [10, 19], the averages of categorical data generally range from 1 to 5 categories, depending on the characteristic. In addition, the average production was quantified.

Obtaining SNPs using the DArTseq™ methodology

We collected samples of young, healthy leaves from everyone at the two Incaper farms (FEM and FEBN) between March and April 2023. The samples were identified, placed in paper bags and kept in a Styrofoam box, with ice packs and thermal blankets. A water sprayer was used to maintain the moisture of the environment. The samples were transported to the Laboratory of Genetics and Plant Breeding of Federal University of Espírito Santo (UFES), Alegre campus. In the laboratory, the samples were stored at -80°C. Subsequently, they were freeze-dried for three days and then stored at -20°C until DNA extraction.

The DNA extraction and purification from the samples followed the CTAB protocol [26], with specific adaptations for coffee [25]. This method uses chloroform, isoamyl alcohol and ethanol (Merk). DNA concentrations and integrity were assessed using a NanoDrop™ 2000 spectrophotometer, and DNA quality was verified by 0.8% agarose gel electrophoresis. Samples that met the quality criteria were forwarded to the Genetic Analysis Service for Agriculture (SAGA), part of the International Maize and Wheat Improvement Center (CIMMYT) in Mexico. The quality criteria required at least 50 µL of DNA dissolved in Tris-EDTA (TE) buffer and suitability for digestion and ligation with restriction enzymes. At SAGA, the samples were subjected to genotyping by sequencing using the DArTseq™ technique [27, 28].

According to the DArTseq™ method, the genome complexity of each sample was reduced using the restriction enzymes *Hpa*II (frequent cutter) and *Pst*I (rare cutter). The ends of the cleaved fragments were ligated to a code adapter and a common sample adapter for individual identification (barcode). The fragments were then subjected to PCR amplification [27]. Subsequently, equimolar amounts of amplicons from each sample were combined in a 96-well microplate, purified and quantified for sequencing using an Illumina NovaSeq 6000 system at the SAGA Laboratory. Barcode sequences of the samples were identified and used to label everyone. Low-quality sequences were filtered and identical sequences were grouped into fastq files. These files were subsequently processed by the DArT PL pipeline software (DArTsoft-seq14) and designated SNPs [28]. The identified SNPs were used in subsequent analyses based on variant calling.

Data filtering and visualization of *C. canephora* chromosomes

We identified a total of 12,418 SNPs in the genotypes by the DArTseq™ method, which were subjected to filtering tests using the "dartR" package v2.9.9.5 [29], in the R software, version 4.2.1 [30], by varying the Reproducibility, Call Rate and MAF parameters. After multiple tests, we decided to use the Reproducibility parameters of 0.98, Call Rate of 0.95 and MAF of 0.05, retaining 1,551 SNPs for subsequent analyses. The subgroups of Conilon, Robusta and hybrid individuals were also analyzed separately. In this case, the filters used were Reproducibility (0.98), Call Rate (0.95) and MAF (0.01), given the small size of the isolated groups. SNP positions were plotted across the 11 chromosomes of the *C. canephora* reference genome (GCA_036785865.1) [31] using CMplot v4.5.1 [32].

Genetic diversity and population structure

The distance between genotypes and clustering analyses were determined by a genetic distance matrix, using Roger's distance and the UPGMA clustering method, with the "genepop" v1.2.11 [33] and "poppr" packages v.2.9.7 [34]. Principal component analysis (PCA) was performed using the "dartR" package v2.9.9.5 [29], based on Euclidean distance. Then, to analyze the population structure of the groups, we performed simulations based

on the Bayesian clustering method, implemented with the nonnegative matrix factorization (sNMF) technique, implemented in the "LEA" package v2.8.0 [35], for each value of K, from 1 to 5, totaling 2,000 iterations. The ancestral proportion of each simulated individual was obtained from the run with the lowest cross-entropy. This analysis allowed estimating the ancestry coefficients and evaluating the relationships between populations. These combined approaches provided a comprehensive understanding of the genetic diversity and population structure of the dataset.

From the defined groups, the expected heterozygosity (He), observed heterozygosity (Ho), inbreeding coefficient (F_{is}) and polymorphism information content (PIC) were estimated using the "HardyWeinberg" package v1.7.9 [36]. The fixation index (F_{st}) was used to evaluate the genetic differentiation between the groups identified in the cluster analysis. Molecular analysis of variance (AMOVA) [37] was conducted using the "poppr" package v2.9.7 [34] to decompose the genetic variation between and within the groups defined by the cluster analysis. The differentiation between the groups and the genetic variation were calculated using the "hierfstat" package v0.5.11 [38].

To observe the genomic profile of the SNPs obtained for the 121 individuals, a heatmap was generated using the heatmap function of the "pheatmap" package v1.0.12 [39]. After these analyses, we verified which individuals maintained the classification when comparing the morphological data and the genetic data found here. For this new dataset, we performed analyses of genetic diversity and population structure. Comparing the botanical groups (Conilon and Robusta), AMOVA was performed. In addition, analyses of genomic profile, distance matrix and structuring of these two groups were determined separately.

SNP discriminant annotation

From the genomic profile of Conilon and Robusta individuals that maintained their classification, both for morphological and genetic data, SNPs with discriminating genotypes of groups were selected. The genomic region of each discriminating SNP in the *C. canephora* reference genome was identified using the Browse Genome tool of the National Center for Biotechnology Information (NCBI) website (https://www.ncbi.nlm.nih.gov), using the genome of *C. canephora* (GCA_036785865.1) [31]. The positions provided by the DArTseq™ method and the chromosomes of the SNPs were entered into the search field of the tool, and we observed whether the SNPs occurred in or near genic regions. The distribution of the discriminating SNPs along the chromosomes of *C. canephora* was verified using the online map Gene2Chrom web v2.1 [40]. All sequencing, chromosome, position and annotation information of the discriminating SNPs were made available.

Validation of discriminating SNPs in a large and mixed population

We carried out a new analysis, including the 121 individuals used and another 529 individuals, forming a single database of 650 *C. canephora* individuals, with the objective of validating the power of the previously selected SNPs for discrimination of the genetic groups corresponding to the botanical groups Conilon and Robusta. Among these 650 individuals, 448 were genotypes from Espírito Santo, originating from seeds collected from parents in the south of Espírito Santo, in the municipalities of Alegre, Cachoeiro de Itapemirim, Jerônimo Monteiro and São José do Calçado. Complementing this set were five commercial clones from nurseries, 56 commercial clones and 111 genotypes from the Incaper germplasm bank, as well as 13 commercial clones and 17 genotypes belonging to the germplasm bank of UFES [25]. This sample covered groups previously characterized morphologically as Conilon, Robusta or Bukobensis by Incaper, in addition to individuals with intermediate traits between Robusta and Conilon.

Results

Genetic Diversity

After filtering, 1,551 SNPs were retained for downstream analyses. PIC ranged from 0.05 to 0.50, with a mean of 0.39 (variance = 0.01). SNPs were distributed across the 11 chromosomes of *C. canephora* (GCA_036785865.1), with greater saturation toward chromosome ends [31]. The number of SNPs per chromosome ranged from 53 to 212, and chromosomes 1, 2, 6, and 7 contained the highest SNP counts (Fig. 1).

Fig. 1. Distribution of 1,551 SNPs across the 11 *C. canephora* chromosomes (GCA_036785865.1) used in the analysis. The x-axis represents chromosome length in megabases (Mb). Colors represent relative SNP marker density along each chromosome, with warmer colors indicating higher density and cooler colors indicating lower density.

There were no duplicate genotypes, and the 121 individuals were arranged in three groups (G1, G2 and G3) (Fig. 2a). Group G1 was composed of 38 individuals (31 Robusta and seven morphological hybrids). Considering the composition, G1 was called Robusta because it grouped 31 of the 33 individuals morphologically described for this group. This group presented the greatest distance in relation to the others and the greatest divergence among its individuals. The Conilon individuals were distributed into two groups: G2 and G3. Group G2 was composed of 12 individuals, of which seven were commercial (six Conilon: V07, BAG110, BAG138, R5, R2, BAG328; two Robusta: IAC51 and BAG558; and four morphological hybrids: A1, 203, V03 and 207). In this group, a genomic profile of the individuals was observed, suggesting they were hybrids between the two botanical groups (Fig. 3). Group G3 was formed by most individuals (71), 46 of which were Conilon and 25 were morphological hybrids. In this group, the lowest genetic divergence occurred between the individuals. This group was classified as Conilon.

The first two principal components explained 87.2% of the existing variation, and three groups were detected. Most of the Conilon and Robusta individuals were grouped in opposite quadrants in the dispersion graph. Most Conilon individuals were grouped in a compact manner, demonstrating greater genetic similarity among individuals, while Robusta individuals showed wider dispersion. Morphological intermediates, in general, showed greater genetic proximity to the Conilon group (Fig. S1).

The population structure analysis revealed three genetic groups (K = 3). The Robusta (G1) and Conilon (G3) groups were well structured, although G1 showed an ancestral proportion shared with G2 and G3, and the G2 (hybrids) group contained half of the gene cluster of G3. In the graphic visualization, the colors used to represent the three clusters are blue for the gene cluster with the highest proportion in G1 individuals, green for the gene cluster with the highest proportion in G2 individuals, and red for the gene cluster with the highest proportion in G3 individuals (Table S1). Each bar in the graph corresponds to an individual evaluated, with the colors related to their respective clusters (Fig. 2b).

Fig. 2. a) Dendrogram showing the genetic dissimilarity among the 121 *C. canephora* individuals, revealing three groups: G1 (Robusta, blue), G2 (hybrids, green), and G3 (Conilon, red). Commercial clones are marked with an asterisk (*). b) Population structure of the 121 individuals, highlighting three genetic components (G1, blue; G2, green; G3, red). c) Scree plot indicating K = 3 as the best number of genetic groups.

The analysis of genetic diversity, using the inbreeding coefficient (F_{is}), Ho and He parameters estimated by groups (Table 3), revealed: lower genetic diversity and higher inbreeding in G3 - Conilon; higher diversity and excess heterozygosity in G2 - hybrid (composed mostly of Conilon, morphological hybrids and Robusta). The

higher values of Ho and He indicate that the individuals in this group have potential to be a source of genetic and genotypic variability. In G1 - Robusta, despite the second highest average value of He (0.21), the F_{is} value (0.16) indicates an excess of homozygotes, even with the general occurrence of only one homozygote (Fig. 3). Furthermore, the genomic profile of Robusta showed only two genotypes per locus, with a higher proportion of heterozygotes. The joint analysis of the 121 individuals revealed that the highest He was 0.34, the lowest Ho was 0.14, and the greatest excess of homozygotes (F_{is}) was 0.59.

Table 3. Average estimates of heterozygosity (Ho and He) and inbreeding coefficient (F_{is}) calculated for the 121 individuals, considering a total of 1,551 SNPs, distributed among three identified genetic groups: Group 1 (G1), composed of 38 individuals and 1,469 SNPs; Group 2 (G2), composed of 12 individuals and 1,422 SNPs; and Group 3 (G3), composed of 71 individuals and 1,343 SNPs.

Genetic group	N°	Botanical group	Polymorphic loci	Ho ± SD	He ± SD	F_{is}
G1	38	Robusta	1,469	0.21 ± 0.16	0.25 ± 0.17	0.16
G2	12	Hybrids	1,422	0.33 ± 0.21	0.31 ± 0.15	- 0.06
G3	71	Conilon	1,343	0.06 ± 0.12	0.08 ± 0.13	0.23
All	121	-	1,551	0.14 ± 0.08	0.34 ± 0.12	0.59

Ho: Mean observed heterozygosity of the group; SD: Standard deviation of the observed heterozygosity of the group; He: Mean expected heterozygosity of the group; SD: Standard deviation of the expected heterozygosity of the group; F_{is}: Coefficient of inbreeding of the group (significant F_{is}).

Considering the groups formed in the cluster analysis (Fig. 2a), we estimated that 78.49% of the genetic variation could be attributed to differentiation between groups and 21.51% within groups. The Phi-samples-total statistics, with a value of 0.78, confirmed this genetic differentiation between the groups. This distribution of variation indicated significant genetic structuring between the groups. Population differentiation (F_{st}) revealed greater genetic differentiation between G1 – Robusta and G3 – Conilon (F_{st} = 0.77), followed by differentiation between G1 – Robusta and G2 – hybrid (F_{st} = 0.46). The genetically closest groups were G2 and G3 (F_{st} of 0.31), composed of Conilon individuals.

In the genomic profile of the identified genetic groups, the 121 individuals are represented by horizontal lines perpendicular to the X axis, while the 1,551 SNPs are grouped vertically according to their genotypes in the individuals (Fig. 3). The Conilon group presented many SNPs in homozygosity, identified by the colors blue (homozygous for the reference allele) and red (homozygous for the alternative allele). The identification of groups of SNPs with homozygous genotypes is explained by the low genotypic variability found in the diversity estimate analysis. In addition, the presence of a small group of SNPs based on wide genotypic variability in the Conilon individuals stands out.

Robusta individuals exhibited a predominance of heterozygous SNPs (yellow), while the same SNPs were predominantly homozygous in Conilon individuals, evidencing that at these loci, polymorphisms occurred only in the Robusta group, which may be a source of genetic variation in crosses with Conilon individuals. A group of potential SNPs was identified to discriminate against the two botanical groups. These SNPs presented homozygous genotypes for the alternative allele (red) in Conilon and homozygous for the reference allele (blue) in Robusta. In group G2 (hybrid), composed of 12 individuals, wide genotypic variability was observed in most SNPs, with the three possible genotypes being detected per SNP. The genomic profiles of these individuals suggested they were genetic hybrids resulting from the cross between individuals of the Robusta and Conilon groups.

Fig. 3. Heatmap of the genomic profiles of 121 *Coffea canephora* individuals classified by Incaper based on

morphology as Conilon (n = 52), Robusta (n = 33), and morphological intermediates (n = 36). The genotype matrix includes 1,551 SNPs. Genotypes are color-coded as follows: 0 (blue) = homozygous reference allele; 1 (yellow) = heterozygous; 2 (red) = homozygous alternative allele; NA (white) = missing data.

Considering these results, a new diversity analysis was performed using only the individuals that maintained the same classification within the botanical groups in both the morphological and genetic analyses (Table 4). Thus, 31 Robusta individuals, 46 Conilon individuals, and 12 probable hybrids detected in this work were analyzed. For these groups, SNPs were re-filtered separately to verify group-specific polymorphisms and to eliminate between-group polymorphisms identified in the previous analyses. In total, 1,925 SNPs were analyzed.

Table 4. Average estimates of heterozygosity and F_{is} calculated for 89 individuals, considering a total of 1,925 SNPs, distributed as follows: Conilon, composed of 46 individuals and 806 SNPs; Robusta, composed of 31 individuals and 1,731 SNPs; and Hybrids, composed of 12 individuals and 1,494 SNPs.

Group	N°	Polymorphic loci	Ho ± SD	He ± SD	F_{is}
Robusta	31	1,731	0.18 ± 0.16	0.21 ± 0.16	0.15
Hybrid	12	1,494	0.26 ± 0.22	0.24 ± 0.17	- 0.06
Conilon	46	806	0.05 ± 0.11	0.05 ± 0.12	0.09

Ho: Mean observed heterozygosity of the group; SD: Standard deviation of the observed heterozygosity of the group; He: Mean expected heterozygosity of the group; SD: Standard deviation of the expected heterozygosity of the group; F_{is}: Coefficient of inbreeding of the group (significant F_{is}).

Although the number of SNPs varied between groups in this analysis, the results were like those obtained previously (Table 3), with greater genetic diversity and excess of heterozygotes in the hybrid group, followed by the Robusta group, which presented Ho and He values of 0.18 and 0.21, respectively, but with a higher F_{is} value. The Conilon group had the lowest genetic diversity, but there was also a reduction in F_{is} in this analysis. The genomic profiles of the Conilon and Robusta groups were analyzed separately for comparison purposes (Fig. 4). A heatmap was generated to show the genomic profile of 46 Conilon individuals, based on 806 SNPs (Fig. 4a), and of 31 Robusta individuals, based on 1,731 SNPs (Fig. 4b). The genetic distance matrix revealed two subgroups, in both Conilon (Fig. 4c) and Robusta (Fig. 4d). Structural analysis showed three gene clusters in both groups, indicating that Conilon presents genetic mixing (Fig. 4e), while Robusta is moderately structured (Fig. 4f).

Fig. 4. (a) Heatmaps of the genomic profiles of 46 Conilon individuals (806 SNPs) and 31 Robusta individuals (1,731 SNPs), classified based on concordant morphological and genetic criteria. Genotypes are color-coded as follows: 0 (blue) = homozygous reference allele; 1 (yellow) = heterozygous; 2 (red) = homozygous alternative allele; NA (white) = missing data. (b) UPGMA dendrogram based on Euclidean distances, showing three clusters across individuals from both botanical groups. (c) Scree plot for K = 3 showing the percentage of variance explained, supporting population structure between the Conilon and Robusta groups.

Annotation and Validation of Discriminating SNPs

Based on the genomic profile of the 121 individuals (Fig. 3), we selected 29 SNPs for molecular discrimination of the genetic groups (Fig. 5a). These SNPs were distributed in nine of the 11 *C. canephora* chromosomes and were located outside the genic regions (Fig. 5b). These markers were then validated in a new dataset, composed of 650 samples (Fig. 5c). These samples included genotypes of a breeding population from seeds of parents collected in southern Espírito Santo (448 individuals), originating from the municipalities of Alegre, Cachoeiro de Itapemirim, Jerônimo Monteiro and São José do Calçado. Complementing this set were five

commercial clones from nurseries, 56 commercial clones and 111 genotypes from the Incaper germplasm bank, in addition to 13 commercial clones and 17 genotypes from the UFES germplasm bank [25]. This broad sample encompassed groups previously characterized morphologically as Conilon, Robusta or Bukobensis, as well as individuals with intermediate traits between Robusta and Conilon. The analysis revealed that 10 of the 29 SNPs were sufficient to discriminate against the three major genetic groups. Furthermore, the population originating from parents in the south of the state presented greater diversity, with a significant number of alleles from both the Conilon and Robusta groups, indicating possible gene flow between the groups in these populations.

Fig. 5. (a) Heatmap of the genomic profiles of 77 *Coffea canephora* individuals classified as Conilon (n = 46) and Robusta (n = 31) based on concordant morphological and genetic criteria, using 29 discriminant SNPs. Genotypes are color-coded as follows: 0 (blue) = homozygous reference allele; 1 (yellow) = heterozygous; 2 (red) = homozygous alternative allele; NA (white) = missing data. (b) Genomic distribution of the 29 discriminant SNPs across 9 of the 11 *C. canephora* chromosomes. (c) Heatmap of an independent validation panel (n = 650) genotyped with the same 29 discriminant SNPs.

Discussion

Genetic diversity and differentiation

In this study, 121 *Coffea canephora* individuals, previously classified as Conilon, Robusta, and hybrids, were assigned to three genetic groups based on SNP analysis. Genetic diversity and population structure were assessed, revealing lower diversity in the Conilon group and higher diversity in a hybrid subgroup within Conilon. Additionally, SNPs discriminating between Conilon and Robusta were identified.

Two genetic groups corresponded to Conilon (G2 and G3) and one to Robusta (G1). Among the morphologically classified hybrids, 64% clustered with G3 (Conilon), 25% with G1 (Robusta), and only 11% with the intermediate group G2, which was genetically closer to Conilon and therefore classified as hybrid. These results indicate that morphological classification based on 29 traits [10, 19] effectively discriminated against Conilon and Robusta but had limited accuracy for identifying intermediate individuals. Similar patterns, in which putative hybrids cluster with Conilon or Robusta, have been reported using microsatellites [11, 41] and SNP markers [2]. Accurate identification of these materials is essential for coffee breeding programs to properly guide crosses and exploit heterosis [42].

One relevant aspect concerns the grouping of morphological hybrids, mostly associated with the Conilon group. This pattern can be explained by two non-mutually exclusive hypotheses: (i) dominance of alleles related to Conilon phenotypes in the evaluated traits; and/or (ii) asymmetry in allelic frequencies resulting from crosses between predominantly heterozygous Robusta individuals and predominantly homozygous Conilon individuals, resulting in hybrids that are genetically closer to Conilon (approximately 0.75 for the Conilon allele and 0.25 for the Robusta allele) and, consequently, also phenotypically similar.

In a recent study on *C. canephora* (Robusta), local ancestry was mapped using the ELAI tool (Efficient Local Ancestry Inference), enabling high-resolution identification of wild genomic segments [43]. The results demonstrated that the distribution of genetic segments varied among chromosomes: some presented completely heterozygous segments, while others presented admixture segments concentrated in specific regions, especially at the terminal ends of chromosomes 4, 9, and 10. These patterns reflect Robusta complex hybridization history, shaped by recombination, backcrossing, and multiple admixture events [43].

The cluster analysis also showed that Robusta (G1) and Conilon (G3) were the most divergent groups (F_{st} = 0.77). This was the highest value reported so far for *C. canephora* in studies with SNPs [2, 44, 6, 45, 23, 25].

G3, formed by the majority of Conilon individuals, presented low genetic diversity among individuals and greater inbreeding, while in G1 (Robusta), there was greater variation among individuals and greater diversity (He) than in the Conilon group. These results indicate the need to increase genetic variation in the Conilon botanical group in Espírito Santo, especially considering that this group represents most commercial clones in the state [10, 19] and that the state is one of the main producers of the species in the world [18].

The low genetic diversity of Conilon compared to Robusta has already been reported in comparative studies [2]. However, the group of genotypes studied here had even lower diversity than the values reported in other studies [41, 25], which may be the cause of the high F_{st} count between the groups. Furthermore, in Espírito Santo, the adoption of clonal cultures has reduced the number of old seed cultures in the state, which maintained segregating populations that may still be valuable genetic resources and sources of alleles for germplasm banks.

Identification and classification of hybrids

The G2 (hybrid) group showed the highest genetic diversity, excess heterozygosity, and admixture of the Conilon and Robusta genetic pools, supporting the hypothesis that these individuals result from crosses between these groups. Morphologically classified hybrids did not form a distinct genetic group, exhibiting genetic profiles distributed among Conilon, Robusta, and intermediate groups, with greater similarity to Conilon, indicating that morphological classification does not exclusively reflect genetic background and may be influenced by environmental factors [46].

The heterozygosities were higher in G2 (with He = 0.31 and Ho = 0.33), which although composed of 12 individuals morphologically classified into different groups — six Conilon (V07, BAG110, BAG138, R5, R2, BAG328), two Robusta (IAC51 and BAG558) and four morphological hybrids (A1, 203, V03 and 207) —, presented a similar genomic profile, with high genotypic variability. This was reflected in the negative inbreeding coefficient (F_{is}), indicating an excess of heterozygotes [45, 47]. This excess of heterozygotes is expected in negative preferential crosses, common in species with self-incompatibility such as *C. canephora* (Musoli *et al*., 2009) and has been reported in groups of *C. canephora* genotypes in different works [16, 41].

High Ho and He values have also been reported for *C. canephora* populations from Africa, with heterozygosity values ranging from approximately 0.30 to 0.41 across different origins and sampling strategies [6, 45, 48]. In Brazil, heterozygosity values like those observed in G2 and G1 have been reported for cultivated *C. canephora* populations [42, 25], whereas lower heterozygosity, comparable to that observed in G3, has been reported for wild or more genetically structured populations [23, 49]. In this study, although group G3 exhibited low heterozygosity values, the remaining groups (Conilon, Robusta, and hybrids) showed high genetic diversity, highlighting the importance of these genotypes for the conservation of *C. canephora* genetic resources.

When analyzed together, the group results showed lower genetic diversity, higher inbreeding and higher population differentiation (F_{st}). Therefore, the joint analysis revealed clear differentiation between the Robusta and Conilon groups. In the separate analysis, based on genotypes that presented the same morphological and genetic characterization, a significant reduction in the number of polymorphic loci was observed in Conilon, confirming its lower genetic variability. The Robusta group, on the other hand, exhibited high genetic diversity and positive F_{is}, indicating a higher proportion of heterozygotes than expected.

In the AMOVA, considering the groups formed in the joint analysis, 78.49% of the genetic variation was attributed to differentiation between groups and 21.51% within groups. In a similar study, Zaidan *et al*. [25] found that for two hierarchical levels (structure and cluster analysis), differences between groups were responsible for 31.88% and 17.76% of genetic variation.

Genetic differentiation (F_{st}) between groups G1 and G3 (F_{st} = 0.77) was higher than that of groups G1 and G2 (F_{st} = 0.46) and groups 2 and 3 (F_{st} = 0.31). These genetic differentiation values were higher than those detected by Anagbogu *et al*. [44], who found for *C. canephora* values between 0.13 and 0.25, and by Akpertey *et al*. [6], who detected a value of F_{st} = 0.26. Similarly, Depecker *et al*. [45] found lower values of F_{st} (0.017 and 0.025), and Verlensey *et al*. [48] also detected a lower value of F_{st} = 0.14. However, other studies have reported values like those of this work, such as Zaidan *et al*. [25], in which the highest value of genetic differentiation detected was F_{st} = 0.60, and Vi *et al*. [43], who studied 65 *C. canephora* genotypes (African and Vietnamese) and found F_{st} = 0.55.

The inbreeding coefficient (F_{is}), which ranges from -1 to +1, measures the balance between homozygotes and heterozygotes, with negative values indicating an excess of heterozygotes [50, 42]. In this study F_{is} values ranged from -0.06 to +0.23, with the negative value observed in group G2, composed of Conilon, Robusta, and hybrid genotypes. Similar negative F_{is} values have been reported in *C. canephora* populations [45, 49, 25], consistent with the outcrossing mating system and gametophytic self-incompatibility of the species [51].

Implications for genetic improvement and conservation

Morphological classification does not always reflect molecular classification, as morphological descriptors may be influenced by environmental factors and dominance effects that mask genetic variation. Accordingly, morphological variation can obscure true genetic diversity [52]. Therefore, the use of molecular descriptors is essential for accurate classification and reclassification of genetic groups, such as G2. To identify the Conilon and Robusta genetic groups, only individuals with concordant morphological and molecular classifications were used to select a minimal set of discriminant SNPs (29). Validation in independent samples showed efficient discrimination among 650 individuals, with only 10 of the 29 SNPs being sufficient to differentiate Robusta, Conilon, and hybrid genotypes.

These results reinforce the importance of integrating molecular data for accurate classification, corroborating previous studies that highlight the value of molecular markers in revealing the true genetic diversity of *C. canephora* [2, 22, 53, 54]. Based on the integrated analysis of morphological and molecular data, Conilon and Robusta were confirmed as distinct genetic groups. The identified discriminant SNPs enabled reliable group discrimination, demonstrating that molecular descriptors provide an effective tool for accurate reclassification of *C. canephora* genetic groups.

Study limitations and future perspectives

Some limitations should be considered. The relatively small number of individuals in the hybrid group may limit the accuracy of genetic diversity estimates for this group. In addition, hybrid identification was based primarily on genomic data, without phenotypic validation under controlled conditions. Future studies integrating genomic, agronomic, and phenotypic information, along with increased sample size and broader geographic representation, will be important to improve the understanding of genotype–phenotype relationships and to refine hybrid classification in *C. canephora*.

Conclusion

This study identified 10 SNP markers that discriminate between the Conilon and Robusta groups of *Coffea canephora*, increasing the accuracy in germplasm classification and supporting breeding programs aimed at hybrid vigor and germplasm conservation. The results highlight the genetic diversity and population structure of these groups, emphasizing the relevance of hybrids. The proposed genetic distance matrix guides cross, and our findings

are fundamental for breeding strategies in *C. canephora* cultivation in Brazil.

Acknowledgments: We thank the Office to Coordinate Improvement of Higher Education Personnel (CAPES); the National Council for Scientific and Technological Development (CNPq); and the Espírito Santo Research and Innovation Support Foundation (FAPES) for funding; along with the Capixaba Institute for Technical Assistance and Rural Extension (Incaper) for providing the genotypes from the Active Germplasm Bank. We also thank the Postgraduate Program in Genetics and Breeding (PPGGM) and the Genetics and Plant Breeding and Biometrics Laboratories for the technical and scientific support during the development of this study.

Declarations:

Financing: This work was carried out with the support of the Coordination for the Improvement of Higher Education Personnel (CAPES, Brasília, DF, Brazil – Funding Code 001, scholarship process number 88887.712907/2022-00). Additionally, this study received financial support from the Espírito Santo Research and Innovation Support Foundation (FAPES, Call No. 20/2022 – Research Fee), with approval by the proponent Márcia Flores da Silva Ferreira. The work was also supported by the National Council for Scientific and Technological Development (CNPq, Brasília, DF, Brazil). The author Rosana G. Oliveira received financial support from CAPES (process number 88887.712907/2022-00).

Ethical declaration

Competing interests: We declare no conflicts of interest.

Authors' contributions: R.G.O. designed the study, collected field samples, prepared material for DNA extraction, coordinated sample submission for sequencing, performed the statistical analyses, and drafted the manuscript. F.A.N.A.A. contributed to statistical analyses and to the writing and discussion of the manuscript. I.R.Z. assisted with statistical analyses and contributed to the writing and discussion of the manuscript. J.F.B.S. contributed to genetic material selection, field collections, and statistical analyses. M.A.S. contributed to field sampling, preparation for DNA extraction, and sample submission for sequencing. A.F. supervised the statistical analyses. M.F.S.F. conceived and designed the study, supervised the study, contributed to manuscript writing, and coordinated the research project. All authors reviewed and approved the final manuscript.

Data availability: The data generated in this study are available from the corresponding author upon reasonable request.

Guidelines for Art and Illustrations:

The graphic program used was R Studio.

References

1. Ferrão, M. A. G., Ferrão, R. G., Fonseca, A. F. A. da Verdin Filho, A. C. & Volpi, P. S. Origin, geographical dispersion, taxonomy and genetic diversity of *Coffea canephora*. In *Conilon Coffee* (eds. Ferrão, R. G., Fonseca, A. F. A. da, Ferrão, M. A. G. & de Muner, L. H.) 85–110 (Incaper, 2019).

2. Alkimim, E. R., Caixeta, E. T., Sousa, T. V., Silva, F. L., Sakiyama, N. S. & Zambolim, L. High-throughput targeted genotyping using next-generation sequencing applied in *Coffea canephora* breeding. *Euphytica* **214**, 50; https://doi.org/10.1007/s10681-018-2126-2 (2018).

3. Montagnon, C., Cubry, P. & Leroy, T. Amélioration génétique du caféier *Coffea canephora* Pierre: connaissances acquises, stratégies et perspectives. *Cah. Agric.* **21**, 143–153; https://doi.org/10.1684/agr.2012.0556 (2012).

4. Oliveira, L. N. L. D. *et al.* Selection of *Coffea canephora* parents from the botanical varieties Conilon and Robusta for the production of intervarietal hybrids. *Ciênc. Rural* **48**, e20170444; https://doi.org/10.1590/0103-8478cr20170444 (2018).

5. Santos, A. V. *et al*. Reaction of *Coffea canephora* clones to the root knot nematode, *Meloidogyne incognita*. *Afr. J. Agric. Res.* **12**, 916–922; https://doi.org/10.5897/AJAR2016.11999 (2017).

6. Akpertey, A., Padi, F. K., Meinhardt, L. & Zhang, D. Effectiveness of single nucleotide polymorphism markers in genotyping germplasm collections of *Coffea canephora* using KASP assay. *Front. Plant Sci.* **11**, 612593; https://doi.org/10.3389/fpls.2020.612593 (2021).

7. Leroy, T. *et al.* Genetics of coffee quality. *Braz. J. Plant Physiol.* **18**, 229–242; https://doi.org/10.1590/S1677-04202006000100016 (2006).

8. Moraes, M. S. *et al.* Characterization of gametophytic self-incompatibility of superior clones of *Coffea canephora*. *Genet. Mol. Res.* **17**, gmr16039876; https://doi.org/10.4238/gmr16039876 (2018).

9. Espindula, M. C., Araújo, L. F. B. de, Schmidt, R., Dias, J. R. M. & Rocha, R. B. Early induction of orthotropic shoots in *Coffea canephora*. *Rev. Ceres* **67**, 281–287; https://doi.org/10.1590/0034-737X202067040005 (2020).

10. Ferrão, M. A. G. *et al.* Variabilidade de *Coffea canephora* do banco ativo de germoplasma do Incaper: caracterização dos acessos com base em descritores mínimos. Circular Técnica n. 08-I (Incaper, 2022).

11. Silva, L. F. Estrutura populacional e diversidade genética de *Coffea canephora* detectadas por marcadores moleculares. *Doctoral thesis*, Universidade Federal de Viçosa (2022).

12. Ferrão, R. G. *et al.* Cultivares de cafés Conilon e Robusta. *Inf. Agropec.* **41**, 17–25 (2020).

13. Berthaud, J. Les ressources génétiques pour l'amélioration des caféiers africains diploïdes: évaluation de la richesse génétique des populations sylvestres et de ses mécanismes organisateurs. Conséquences pour l'application. PhD thesis, *Université de Paris* (1984).

14. Cubry, P. *et al.* 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 (2008).

15. Montagnon, C., Leroy, T. & Eskes, A. B. Amélioration variétale de *Coffea canephora*; 1: critères et méthodes de sélection. *Plant. Rech. Dév.* 5, 18–33 (1998).

16. Musoli, P. *et al.* Genetic differentiation of wild and cultivated populations: diversity of *Coffea canephora* Pierre in Uganda. *Genome* **52**, 634–646; https://doi.org/10.1139/G09-037 (2009).

17. Merot-L'anthoene, V. *et al.* Development and evaluation of a genome-wide **coffee** 8.5K SNP array and its application for high-density genetic mapping and for investigating the origin of *Coffea arabica* L. *Plant Biotechnol. J.* **17**, 1418–1430; https://doi.org/10.1111/pbi.13066 (2019).

18. CONAB. *Acompanhamento da safra brasileira de café – Safra 2024.* CONAB (2024).

19. Ferrão, M. A. G. *et al.* Characterization and genetic diversity of *Coffea canephora* accessions in a germplasm bank in Espírito Santo, Brazil. *Crop Breed. Appl. Biotechnol.* **21**, e36132123; https://doi.org/10.1590/1984-70332021v21n2a32 (2021).

20. Brasil. Secretaria de Apoio Rural e Cooperativismo. Serviço Nacional de Proteção de Cultivares. Portaria n. 2, de 17 de novembro de 2000. *Diário Oficial da República Federativa do Brasil* (Brasília, DF), 6–7 (2000).

21. Jamali, S. H., Cockram, J. & Hickey, L. T. Insights into deployment of DNA markers in plant variety protection and registration. *Theor. Appl. Genet.* **132**, 1911–1929; https://doi.org/10.1007/s00122-019-03348-7 (2019).

22. Bikila, B. A., Sakiyama, N. S. & Caixeta, E. T. SNPs based molecular diversity of *Coffea canephora*. *J. Microbiol. Exp.* **5**, 00136; https://doi.org/10.15406/jmen.2017.05.00136 (2017).

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

24. Spinoso-Castillo, J. L. *et al.* Genetic diversity of coffee (*Coffea* spp.) in Mexico evaluated by using DArTseq and SNP markers. *Genet. Resour. Crop Evol.* **67**, 2385–2400; https://doi.org/10.1007/s10722-020-00940-5 (2020).

25. Zaidan, I. R. *et al.* Diversity and structure of *Coffea canephora* from old seminal crops in Espírito Santo, Brazil: genetic resources for coffee breeding. *Tree Genet. Genomes* **19**, 1; https://doi.org/10.1007/s11295-023-01594-x (2023).

26. Doyle, J. J. & Doyle, J. L. Isolation of plant DNA from fresh tissue. *Focus* **12**, 13–15 (1990).

27. Kilian, A. *et al.* Diversity arrays technology: a generic genome profiling technology on open platforms. In *Data Production and Analysis in Population Genomics* (eds. Pompanon, F. & Bonin, A.) *Methods Mol. Biol.* **888**, 67–89 (Humana Press, 2012); https://doi.org/10.1007/978-1-61779-870-2_5.

28. Sansaloni, C. *et al.* Diversity analysis of 80,000 wheat accessions reveals consequences and opportunities of selection footprints. *Nat. Commun.* **11**, 4572; https://doi.org/10.1038/s41467-020-18404-w (2020).

29. Gruber, B., Unmack, P. J., Berry, O. F. & Georges, A. dartr: an R package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Mol. Ecol. Resour.* **18**, 691–699; https://doi.org/10.1111/1755-0998.12745 (2018).

30. R Core Team. *R: a language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna, Austria, 2022); https://doi.org/10.32614/R.manuals.

31. Denoeud, F. *et al.* The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* **345**, 1181–1184; https://doi.org/10.1126/science.1255274 (2014).

32. Yin, L. *et al.* CMplot: circle Manhattan plot. R package version 4.5.1 (2021)

33. Rousset, F. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Mol. Ecol. Resour.* **8**, 103–106; https://doi.org/10.1111/j.1471-8286.2007.01931.x (2008).

34. Kamvar, Z. N., Tabima, J. F. & Grünwald, N. J. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* **2**, e281; https://doi.org/10.7717/peerj.281 (2014).

35. Frichot, E. & François, O. LEA: an R package for landscape and ecological association studies. *Methods Ecol. Evol.* **6**, 925–929; https://doi.org/10.1111/2041-210X.12382 (2015).

36. Graffelman, J. & Weir, B. S. Testing for Hardy-Weinberg equilibrium at bi-allelic genetic markers on the X chromosome. *Heredity* **116**, 558–568; https://doi.org/10.1038/hdy.2016.20 (2016).

37. Excoffier, L., Smouse, P. E. & Quattro, J. M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479–491; https://doi.org/10.1093/genetics/131.2.479 (1992).

38. Goudet, J. HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes* **5**, 184–186; https://doi.org/10.1111/j.1471-8286.2004.00828.x (2005).

39. Kolde, R. pheatmap: pretty heatmaps. R package version 1.0.12; https://doi.org/10.32614/CRAN.package.pheatmap (2019).

40. Chao, J. T. *et al.* MapGene2Chrom, a tool to draw gene physical map based on Perl and SVG languages. *Yi Chuan* **37**, 91–97; https://doi.org/10.16288/j.yczz.2015.01.013 (2015).

41. Souza, F. F. *et al.* Molecular diversity in *Coffea canephora* germplasm conserved and cultivated in Brazil. *Crop Breed. Appl. Biotechnol.* **13**, 221–227; https://doi.org/10.1590/S1984-70332013000400001 (2013).

42. Sousa, P. *et al*. *Coffea canephora*: heterotic crosses indicated by molecular approach. *Plants* **11**, 3023;

https://doi.org/10.3390/plants11223023 (2022).

43. Vi, T. *et al*. Genome-wide admixture mapping identifies wild ancestry-of-origin segments in cultivated Robusta coffee. *Genome Biol. Evol.* **15**, evad065; https://doi.org/10.1093/gbe/evad065 (2023).

44. Anagbogu, C. F. *et al.* 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**, 685–696; https://doi.org/10.1007/s10722-019-00744-2 (2019).

45. Depecker, J. *et al.* Genetic diversity and structure in wild Robusta coffee (*Coffea canephora* A. Froehner) populations in Yangambi (DR Congo) and their relation to forest disturbance. *Heredity* **130**, 145–153; https://doi.org/10.1038/s41437-022-00588-0 (2023).

46. Motta, L. B. *et al.* Transferability of microsatellite loci in *Coffea canephora*. *Aust. J. Crop Sci.* **8**, 987–991 (2014).

47. Moura, T. M. *et al.* Diversidade e estrutura genética espacial em populações fragmentadas de *Solanum* spp. do Cerrado, estimadas por meio de locos microssatélites. *Sci. For.* **37**, 143–150 (2009).

48. Verleysen, L. *et al.* Crop-to-wild gene flow in wild coffee species: the case of *Coffea canephora* in the Democratic Republic of the Congo. *Ann. Bot.* **133**, 917–930; https://doi.org/10.1093/aob/mcae034 (2024).

49. Souza, L. C. *et al.* Molecular characterization of parents and hybrid progenies of conilon coffee. *An. Acad. Bras. Cienc.* **93**, e20201649; https://doi.org/10.1590/0001-3765202120201649 (2021).

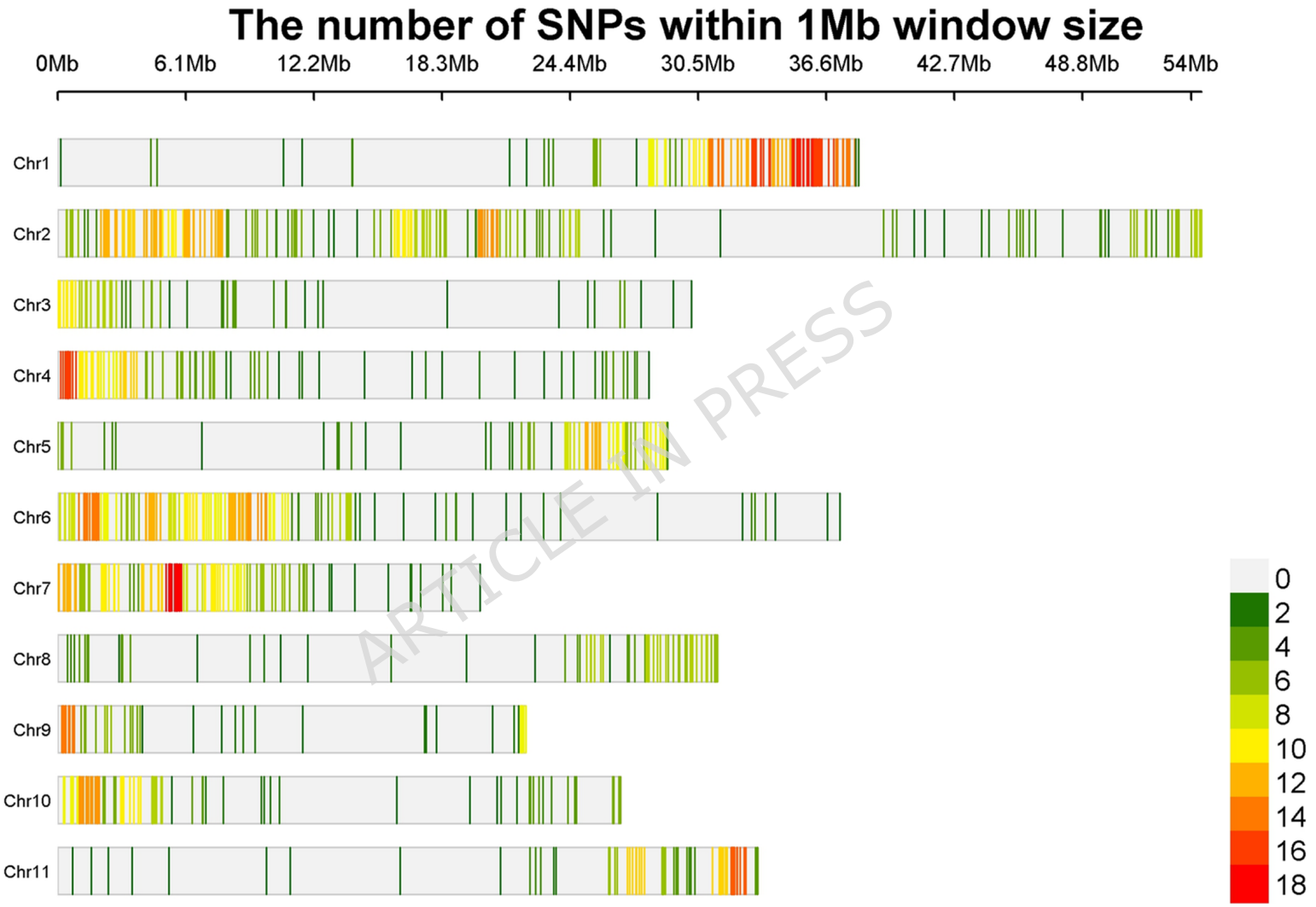
50. Peakall, R & Smouse, P. E. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics* **28**, 2537–2539; https://doi.org/10.1093/bioinformatics/bts460 (2012).

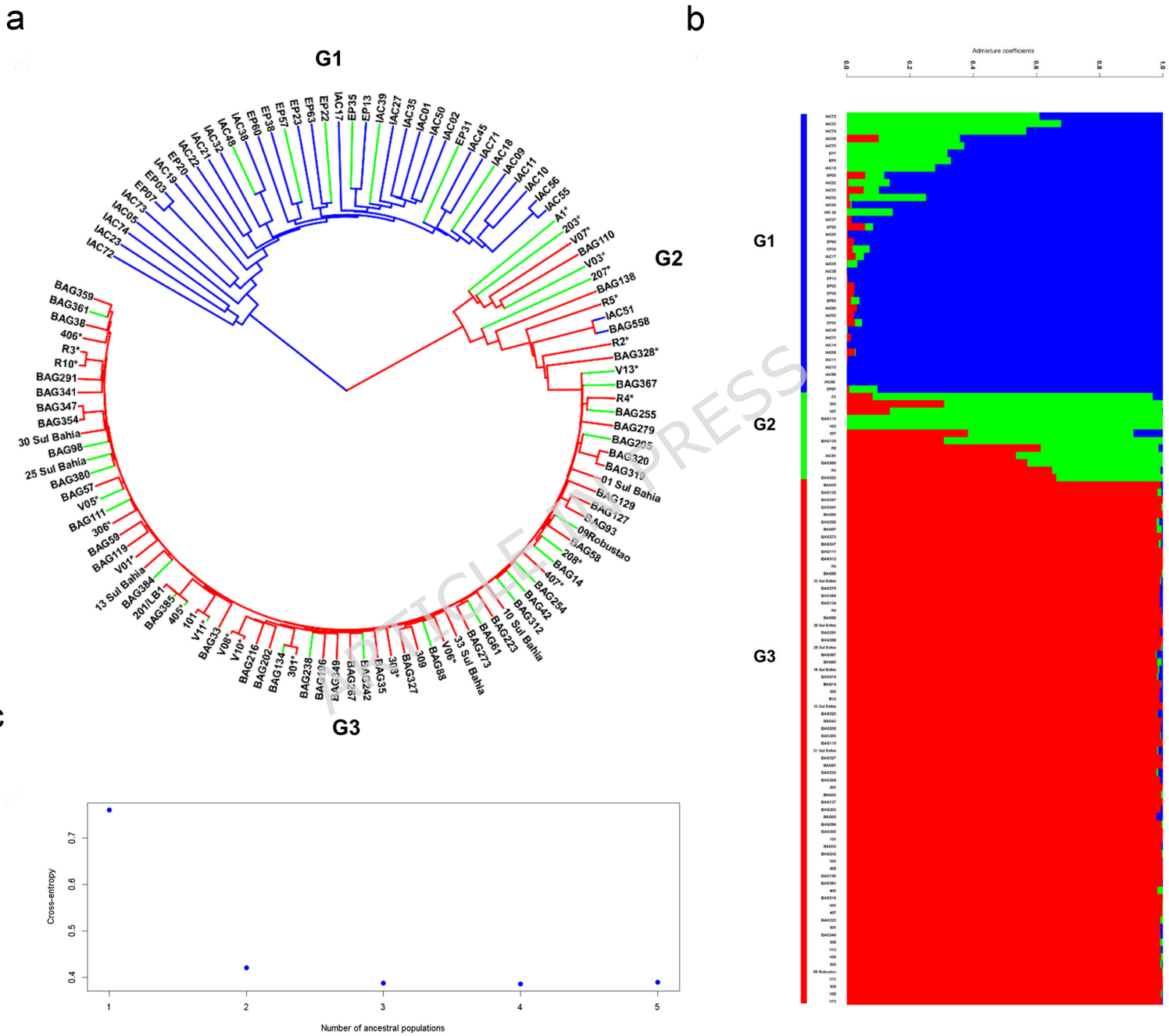
51. Mateu-Andrés, I. & de Paco, L. Genetic diversity and the reproductive system in related species of *Antirrhinum*. *Ann. Bot.* **98**, 1053–1060; https://doi.org/10.1093/aob/mcl186 (2006).

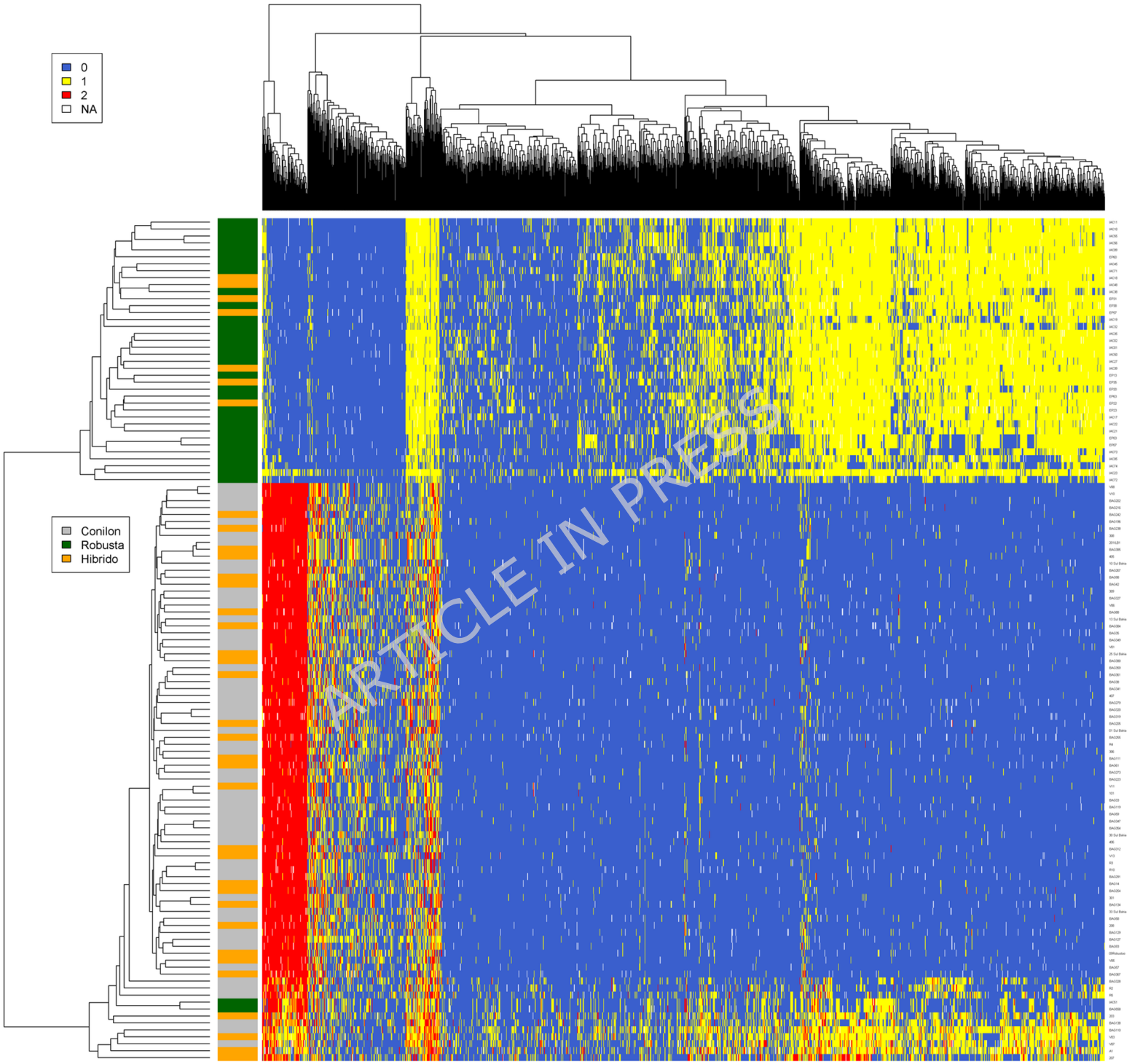
52. Silvestrini, M. *et al.* Genetic diversity of a *Coffea* germplasm collection assessed by RAPD markers. *Genet. Resour. Crop Evol.* **55**, 901–910; https://doi.org/10.1007/s10722-007-9295-5 (2008).

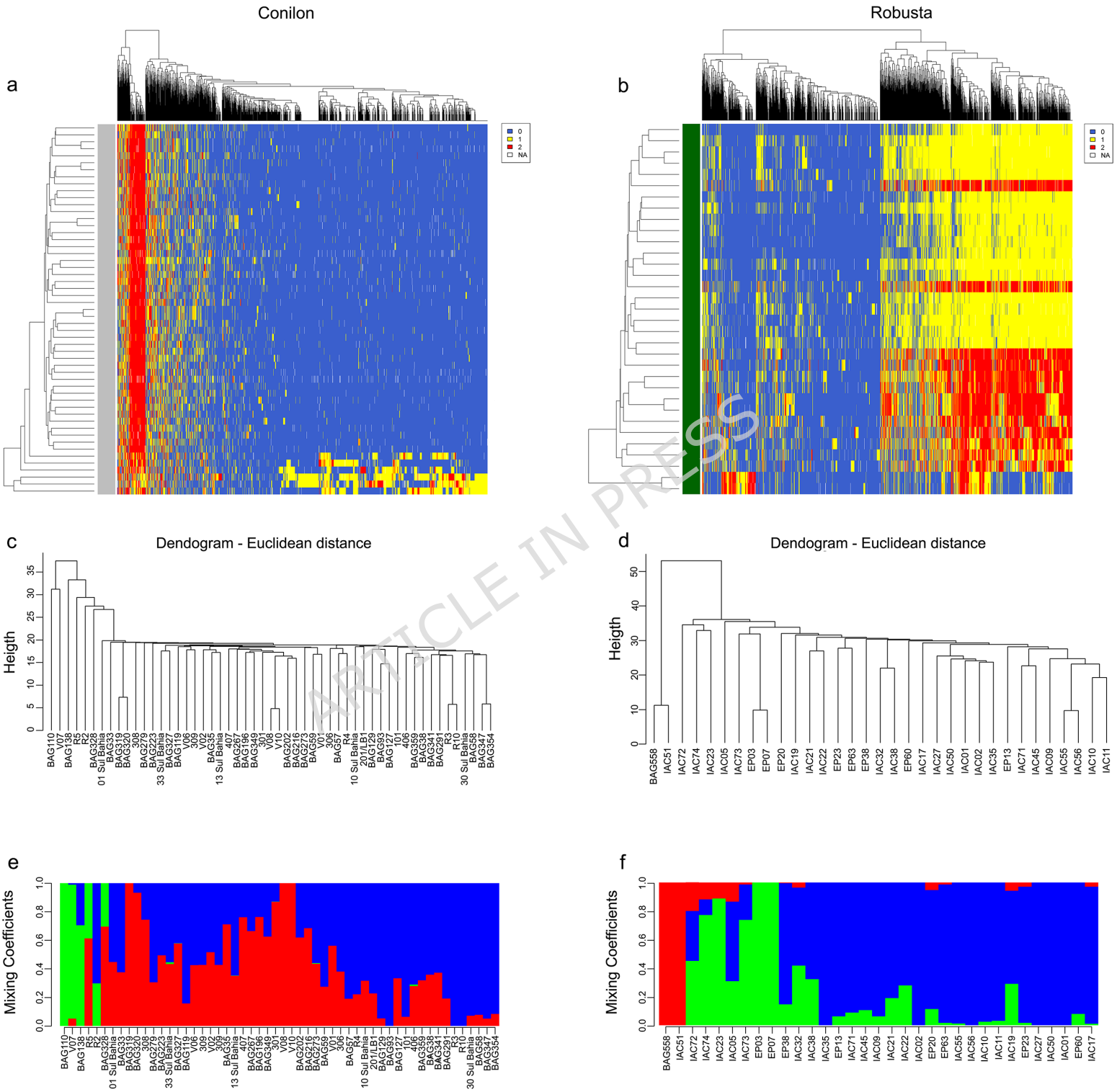
53. Ferrão, L. F. V. *et al.* Comparative study of different molecular markers for classifying and establishing genetic relationships in *Coffea canephora*. *Plant Syst. Evol.* **299**, 225–238; https://doi.org/10.1007/s00606-012-0717-2 (2013).

54. Prakash, N. S., Combes, M. C., Dussert, S., Naveen, S. & Lashermes, P. Analysis of genetic diversity in Indian robusta coffee genepool (*Coffea canephora*) in comparison with a representative core collection using SSRs and AFLPs. *Genet. Resour. Crop Evol.* **52**, 333–343; https://doi.org/10.1007/s10722-003-2125-5 (2005).

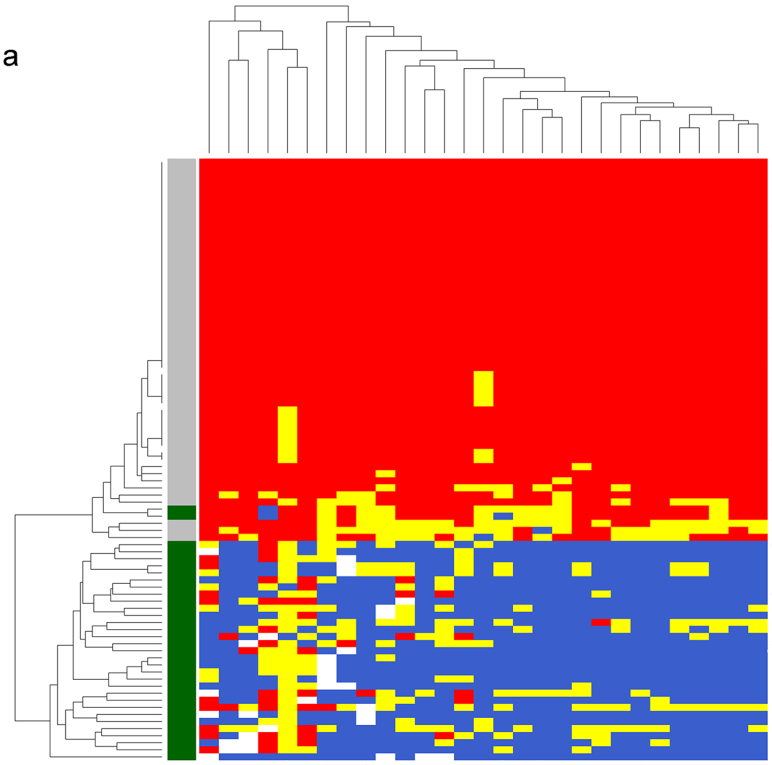




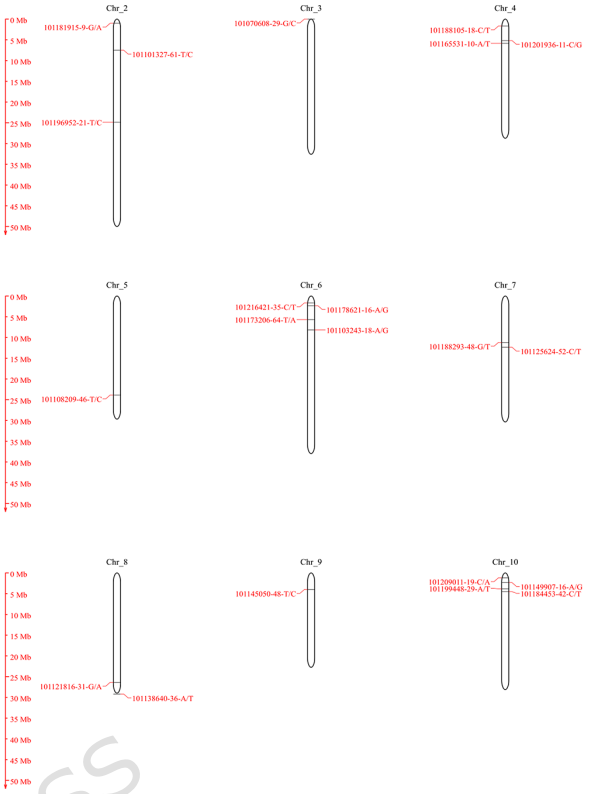




a



b



c