Plant Syst Evol (2014) 300:1649–1661
DOI 10.1007/s00606-014-0990-3

ORIGINAL ARTICLE

The effects of encoding data in diversity studies and the applicability of the weighting index approach for data analysis from different molecular markers

Luís Felipe V. Ferrão · Eveline T. Caixeta · Cosme D. Cruz · Flávio F. de Souza · Maria Amélia G. Ferrão · Eunize Maciel-Zambolim · Laércio Zambolim · Ney S. Sakiyama

Received: 25 September 2013 / Accepted: 15 January 2014 / Published online: 11 February 2014

Abstract The use of molecular markers to study genetic diversity represents a breakthrough in this area, because of the increase in polymorphism levels and phenotypic neutrality. Codominant markers, such as microsatellites (SSR), are sensitive enough to distinguish the heterozygotes in genetic studies. Despite this advantage, there are some studies that ignore this feature and work with encoded data because of the simplicity of the evaluation, existence of polyploids and need for the combined analysis of different types of molecular markers. Thus, our study aims to investigate the consequences of these encodings on simulated and real data. In addition, we suggest an alternative analysis for genetic evaluations using different molecular markers. For the simulated data, we proposed the following two scenarios: the first uses SNP markers, and the second SSR markers. For real data, we used the SSR genotyping data from *Coffea canephora* accessions maintained in the Embrapa Germplasm Collection. The genetic diversity was studied using cluster analysis, the dissimilarity index, and the Bayesian approach implemented in the STRUCTURE software. For the simulated data, we observed a loss of genetic information to the encoded data in both scenarios. The same result was observed in the coffee studies. This loss of information was discussed in the context of a plant-breeding program, and the consequences were weighted to germplasm evaluations and the selection of parents for hybridization. In the studies that involved different types of markers, an alternative to the combined analysis is discussed, where the informativeness, coverage and quality of markers are weighted in the genetic diversity studies.

Keywords Codominant markers · *Coffea canephora* · Dominant markers · Germplasm · SSR · STRUCTURE

Introduction

The accurate evaluation of genetic dissimilarity between genotypes is important in diversity studies. Different methodologies can be used for these evaluations. In the 1960s,

L. F. V. Ferrão · E. Maciel-Zambolim
BIOAGRO, BIOCAFÉ, Universidade Federal de Viçosa, Viçosa, MG 3650-000, Brazil

E. T. Caixeta (✉)
Embrapa Café, BIOAGRO, BIOCAFÉ, Universidade Federal de Viçosa, Viçosa, MG 36570-000, Brazil
e-mail: eveline.caixeta@embrapa.br

C. D. Cruz
Departamento de Biologia Geral, Universidade Federal de Viçosa, Viçosa, MG 36570-000, Brazil

F. F. de Souza
Embrapa Rondônia, BR 364, Km 5,5, Zona Rural. Caixa Postal 127, Porto Velho, RO 76815-000, Brazil

M. A. G. Ferrão
Embrapa Café/Incaper, Rua Afonso Sarlo, 160 Bento Ferreira, Vitória, ES 29052-010, Brazil

L. Zambolim
Departamento de Fitopatologia, Universidade Federal de Viçosa, Viçosa, MG 36570-000, Brazil

N. S. Sakiyama
Departamento de Fitotecnia, Universidade Federal de Viçosa, Viçosa, MG 36570-000, Brazil

genetic diversity was quantified using visual evaluations, such as evaluating the expression of phenotypic markers. With the emergence of molecular biology in the 1980s, these same studies have been carried out using molecular markers. This technology allowed for the detection of polymorphisms at the DNA level and increased the levels of polymorphisms that are able to be accessed. Furthermore, the DNA data present phenotypic neutrality, meaning that the disturbing influences of the environment are not considered in the analysis. These features have made molecular analysis a powerful tool for diversity studies due to higher accuracy and the reliability of the results obtained (Souframanien and Gopalakrishna 2004).

The advances in molecular technologies lead to new perspectives in diversity characterization, and different statistical approaches can be used. Approaches that use allele frequency are based on different parameters to measure the structure and genetic variation presented within populations or a set of genotypes. The comparative perspective approach can be performed among and within a population or groups of individuals. In this case, the genetic dissimilarity (or similarity) matrix is calculated based on the analysis of all possible pairwise combinations of genotypes (Karp et al. 1997; Kosman and Leonard 2005). These results in association with multivariate statistical methods allow the summary, classification and ordering of the information observed (Mohammadi and Prasanna 2003). In addition to these two approaches, a Bayesian technique has been used for studies of genetic diversity and population structure. Implemented in the STRUCTURE software (Pritchard et al. 2000), genotypic data are used for probability classifications of each genotype, taking into account the K populations (where K may be unknown). This approach allows us to obtain robust results and make inferences about migration rate, allele frequency and hybrid zones using dominant and codominant molecular markers.

In plant breeding programs, the use of the comparative approach in germplasm banks is more common in studies of genetic resources (Laurentin 2009). Generally, in this type of study, genetic diversity is evaluated using dissimilarity coefficients to establish the genetic distance matrices. Thus, the use of robust coefficients is a key for the determination of the true genetic variability. The choice of the most appropriate coefficients depends on the type of markers, ploidy of the organism and the objective of each study (Kosman and Leonard 2005). To separate the types of markers, two classes are formed in accordance with the discriminatory ability. The first is formed by dominant markers, which are not able to distinguish the heterozygous genotypes. Included in this class are the following: random amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), inter-simple sequence repeat (ISSR) and diversity arrays technology (DArTs). The other class is composed of codominant markers, which are able to distinguish heterozygous genotypes in the molecular assays. Examples of these types of markers are restriction fragment length polymorphism (RFLP), simple sequence repeats (SSR) and single nucleotide polymorphism (SNP).

The possibility of distinguishing the heterozygous genotypes is an advantage of codominant markers because the diversity analysis is enriched (Ferrão 2013). However, in some studies, researchers choose to encode the molecular data in a binary way rather than discriminating the alleles and use specific coefficients for codominant data. The main reasons for this choice include the easy evaluation in different ploidy levels in some species and the need for combined analysis using dominant and codominant markers. Scoring the alleles in a binary format, which is made possible by keeping a record of only the presence or absence of the bands, simplifies the evaluation and statistical analyses (Kosman and Leonard 2005). Another factor that complicates the codominant marker evaluations is the polyploid analysis. In these evaluations, it is not possible to identify how many allele copies are present in a given heterozygote by visual analysis. Thus, a simple and rapid way is to encode the codominant marker data (Bruvo et al. 2004; De Silva et al. 2005). Moreover, with the availability of different molecular techniques, it is common to use more than one molecular marker type to study genetic diversity (Belaj et al. 2003; Ferrão 2013; Gallego et al. 2005; Lamia et al. 2010). However, each molecular technique differs in informativeness, genome coverage and the quality of data generated. Thus, one way of aggregating information from different methodologies is by evaluating all the markers with the same dissimilarity (or similarity) coefficient and encoding the data.

This study aims to answer the following questions: (1) using simulated data in diploids, what is the loss of information when codominant markers are evaluated as dominant? (2) Using real data, how do these differences in evaluations affect the management of genetic resources in plant breeding programs? (3) What alternative would be best for the joint analysis of data originating from different molecular markers?

Materials and methods

Simulated data

Two different scenarios to study genetic diversity were proposed. In scenario 1, one population of 200 genotypes and 500 biallele loci was simulated. We used SNP markers in this diversity study. In scenario 2, one population with the same sample size and number of loci was simulated. However, we used SSR, a multi-allelic marker (1–9

alleles). The simulation process was conducted using a number of simulation samples called replicas. Each replica (r) was formed using an initial number of markers (m) to be evaluated. A designated increase (Δ) in the number of markers was added to the initial number (m). Thus, it was possible to establish replicas (r) that varied from an initial size (m) to a final size (m′) and an arithmetic ratio (Δ). In both scenarios, the value of m and Δ was 50, while the value of m′ was 500. Therefore, ten replicas (r) were used in each scenario.

In scenario 1, all replicas ranging from m to m′ were evaluated as codominant, and the genetic diversity analysis was performed. In these analyses, each allele received a label according to its molecular size. Afterwards, we used the same replicas (r); however, the markers were encoded as the presence or absence of the band (binary data). In scenario 2, we used the same strategies and another encoding was included for comparison. The most frequent allele in the population was coded as 1, while the other alleles were designated as 0. This transformation is common in studies that use multi-allelic markers.

As a parameter for comparison, we determined that true genetic dissimilarity was obtained by analyzing 500 loci evaluated as codominant. To quantify the informativeness loss caused by data encoding, each replica was evaluated using an appropriate index. For the binary encoded data, genetic dissimilarity was calculated using the complement of Jaccard (1908) coefficient, commonly used in dominant molecular analyses. The codominant data were evaluated using the complement of weighted index (Cruz et al. 2011), cited by Ferrão (2013). The comparison of genetic dissimilarity was performed using the correlation between matrices. The normalization of the Mantel statistic was used to determine the association between two matrices, and 1,000 random permutations were used to test the significance of the matrix correlations.

Simulations and analyses of genetic diversity were performed using the GENES software (Cruz 2013). Five simulations were conducted for each scenario. The results presented are an arithmetic average of the values obtained.

Molecular analysis of *Coffea canephora* access

The genetic diversity studies using real data were performed on the coffee species *C. Canephora* ($2n = 2x = 22$). Eighty-seven accessions from the Embrapa Rondônia Germplasm Bank were used in these analyses (Table 1). Souza (2011) and Ferrão (2013) have already characterized and studied this collection. The accessions that are maintained in this collection belong to two distinct varietal groups: Conilon and Robusta. Natural hybrids between these varietal groups should be considered because *C. canephora* is an allogamous species.

In the laboratory analyses, young and fully developed leaves from each plant were collected, frozen at −80 °C, lyophilized, triturated and stored at −20 °C in the Coffee Biotechnology Laboratory (BioCafé/Bioagro), Brazil. The genomic DNA was extracted according to the protocol described by Diniz et al. (2005), and the molecular analyses were performed using the codominant and multi-allelic SSR marker. Forty-seven SSR primers were used for genotyping (Table 2), and microsatellite amplification was performed as reported by Missio et al. (2010). For the allele score, we used the same methodology proposed in scenario 2 for the simulated data. Thus, the alleles were evaluated as follows: codominant (*Cod*), binary format (*Bin*) or dominant using the most frequent allele (*Dom*).

In the encoded data analyses (*Bin* and *Dom*), the genetic dissimilarity was calculated using the complement of Jaccard (1908) coefficient. For the codominant data (*Cod*), we used the complement of weighted index (Cruz et al. 2011). In all analyses, the dendrograms were constructed using the neighbor joining (NJ) method, and the statistical procedures were performed using the GENES (Cruz 2013) and Figtree v1.3.1 (Rambaut 2006) software.

The Bayesian clustering was performed using the STRUCTURE software (Pritchard et al. 2000). We used K values that ranged from 1 to 5 with mixture models and five repetitions. Each running was implemented with a period of 10,000 burn-in followed by 100,000 MCMC. The number of genetic groups was estimated by the ΔK value (Evanno et al. 2005) using the STRUCTURE HARVESTER software (Earl and vonHoldt 2012).

Results and discussion

Simulated data

For the simulated data, two scenarios were proposed. The first scenarios simulated SNP markers, which are codominant and biallelic. This class of molecular marker is frequently used in genetic studies, because it provides a large amount of information using molecular assays. For these analyses, we observed a difference of 18 % in the correlation value between the coded and unencoded data. These results represent a loss of genetic information when the data are encoding (Fig. 1a).

In the second scenario, we simulated the use of SSR markers, which are multi-allelic and codominant. These features increase the informativeness of the genetic analyses. Two types of encoding were performed. In the first, called *Bin*, we observed a loss of information on the order of 3 %. In these analyses, the codominant and multi-allelic marker was evaluated using the binary format (presence or absence of the band). In the other encoding, called *Dom*, the loss of

Table 1 *Coffea canephora* accessions maintained in the Germplasm Bank of Embrapa Rondônia

Accessions	Code	Accessions	Code	Accessions	Code
Conilon Incaper 03	ES03	Cpafro183	RO183	Cpafro147	RO147
Conilon Incaper 110	ES110A	Cpafro015	RO015	Cpafro164	RO164
Conilon Incaper 28	ES028	Cpafro140	RO140	Cpafro189	RO189
Conilon Incaper 16	ES16	Cpafro001	RO001	Cpafro190	RO190
Conilon Incaper 45	ES45	Cpafro016	RO016	Cpafro036	RO036
Kouillou IAC 661	K661	Cpafro044	RO044	Cpafro089	RO089
Conilon Incaper V.1	ESV1	Cpafro101	RO101	Cpafro045B	RO045B
Conilon Incaper V.2	ESV2	Cpafro119	RO119	Cpafro077	RO077
Conilon Incaper V.3	ESV4	Cpafro155	RO155	Cpafro138	RO138
Robusta IAC 1675	R1675	Cpafro004	RO004	Cpafro142	RO142
Robusta IAC 2259	R2259	Cpafro042	RO042	Cpafro196	RO196
Robusta IAC 2257.1	R22571	Cpafro098	RO098	Cpafro193	RO193
Robusta IAC 2257.2	R22572	Cpafro160	RO160	Cpafro049	RO049
Robusta IAC 640.1	R6401	Cpafro184	RO184	Cpafro030	RO030
Robusta IAC 640.2	R6402	Cpafro018	RO018	Cpafro032	RO032
Robusta IAC 2258.1	R22581	Cpafro045	RO045	Cpafro076	RO076
Robusta IAC 2258.2	R22582	Cpafro146	RO146	Cpafro161	RO161
Robusta IAC 2258.3	R22583	Cpafro194	RO194	Cpafro035	RO035
Cpafro 006	RO006	Cpafro017	RO017	Cpafro038	RO038
Cpafro 047	RO047	Cpafro043	RO043	Cpafro073	RO073
Cpafro 199	RO199	Cpafro120	RO120	Cpafro139	RO139
Cpafro052	RO052	Cpafro010	RO010	Cpafro141	RO141
Cpafro151	RO151	Cpafro064	RO064	Cpafro171	RO171
Cpafro172	RO172	Cpafro086	RO086	Cpafro197	RO197
Cpafro003	RO003	Cpafro103	RO103	Cpafro025	RO025
Cpafro058	RO058	Cpafro203	RO203	Cpafro026	RO026
Cpafro059	RO059	Cpafro022	RO022	Cpafro072	RO072
Cpafro088	RO088	Cpafro024	RO024	Cpafro075	RO075
Cpafro096	RO096	Cpafro127	RO127	Cpafro115	RO115

Adapted from Souza (2011) and Ferrão (2013)

information was 25 %. In this case, those genotypes that presented the most frequent allele were coded as 1 (one), and those that presented other alleles were coded as 0 (zero) (Fig. 1b). For all encoding approaches, the genetic distance between the pairs of genotypes was measured using the Jaccard (1908) coefficient, which is an appropriate methodology for dominant marker studies. Thus, the simulated data indicated that the use of encodings in combination with the statistical methods for dominant marker resulted in a loss of genetic information.

In the genotyping studies using codominant data, it is recommended that the genetic similarities between the pairs of individuals should not be determined based on the proportion of bands that are shared between two individuals, as in the binary data evaluations. The evaluation should be adjusted so that there is a representation of individual allelic patterns across all loci studied, taking into account the total number of loci and the number of shared alleles between the loci (Kosman and Leonard 2005). Comparative studies indicated that the weighted index is an efficient algorithm for determining the diversity between pairs of genotypes because it uses all of the parameters mentioned (Ramos et al. 2011).

Molecular analysis of *Coffea canephora* accessions

The real data were used to quantify the results of the encodings in the genetic studies from the germplasm collection. We used *C. canephora* accessions and cluster analysis based on the genetic dissimilarity coefficients and the Bayesian approach in the STRUCTURE software (Pritchard et al. 2000). The Bayesian analysis allows a probabilistic classification of genotypes into populations according to their ancestry coefficient (*Q*). The SSR data were encoded as *Cod*, *Bin* or *Dom*, as in Scenario 2 of the simulated data. In all analyses (*Cod*, *Bin* and *Dom*), it was possible to classify the access in the respective varietal groups (Conilon or Robusta) in both approaches.

Table 2 Microsatellite primers used in the *Coffea canephora* analysis

Primer	Forward primer (5′ > 3′)	Reverse primer (5′ > 3′)
SSR-07[a]	TGACATAGGGGGCTAAATTG	TTAATGGTGACGCTTTGATG
SSR-08[a]	CACTGGCATTAGAAAGCACC	GGCAAAGTCAATGATGACTC
SSR-13[a]	TGGCCGTGATAATAAACAGC	ATGTGGCAATCTAAAGCCAA
SSR-16[b]	ACCCGAAAGAAAGAACCAAG	CCACACAACTCTCCTCATTC
SSR-21[b]	GACCATTACATTTCACACAC	GCATTTTGTTGCACACTGTA
SSR-29[c]	GGCTTCTTGGGTGTCTGTGT	CCATTGGCTTTGTATTTCTGG
SSR-30[d]	ATGGGGCCAACTTGAATATG	CAGGGCATCTATCTACTTCTCTTT
SSR-34[d]	GGAGACGCAGGTGGTAGAAG	TCGAGAAGTCTTGGGGTGTT
SSR-35[d]	CTGGCATTAGAAAGCACCTTG	GCTTGGCTCACTGTAGGACTG
SSR-37[d]	CAACACTATCTCTTGATTTTTCACT	CGTGCAAGTCACATACTTTACTAC
SSR-39[c]	TCCCCCATCTTTTTCTTTCC	GGGAGTGTTTTTGTGTTGCTT
SSR-40[d]	AAAGGAAAATTGTTGGCTCTGA	TCCACATACATTTCCCAGCA
SSR-42[c]	TTGCTTGCTTGTTCTGTTAT	TGACACGAGAGTTAGAAATGA
SSR-43[d]	TTTTCTGGGTTTTCTGTGTTCTC	TAACTCTCCATTCCCGCATT
SSR-46[d]	AATGAAGAAGAGGGTGGTG	CGAGGGTATTGTTTTCCAG
SSR-48[c]	AGCAAGTGGAGCAGAAGAAG	CGGTGAATAAGTCGCAGTC
SSR-49[d]	TGGAGAAGGCTGTTGAAACC	GGCGTGAAGCAAAAAGGTAT
SSR-52[d]	GACCAAATGTCAGCTCATTG	GCCGACTGCTCTTTTAGTGT
SSR-55[c]	GCAGGTATTTGAAGGATGAACC	GTGTAGGTGGTGCGATGTGT
SSR-56[d]	AGCTATCTTTATCTCACACACACA	GTTAGTGTTTCGATTTGGTACTG
SSR-57[b]	CTCGCTTTCACGCTCTCTCT	CGGTATGTTCCTCGTTCCTC
SSR-59[d]	CCAGCTCTCCTCACTCTTTTCA	GGTGGTGGAGGGGTAATAGG
SSR-64[d]	GTTAGTGTGCGACCGTGTGT	TTATGCCCCTCCCCATATCT
SSR-65[d]	CTCACACACACACTGTCACG	CGAATGAGGCTCCATCAC
SSR-70[e]	GTAACCACCACCTCCTCTGC	TGGAGGTAACGGAAGCTCTG
SSR-71[e]	GCTAAGTTCAATTGCCCCTGT	GGGTTAATTTGATTGCGTGA
SSR-74[e]	TGGGGAAAAGAAGGATATAGACAAGAG	GAGGGGGGCTAAGGGAATAACATA
SSR-76[e]	GGTCCCACTCTCAAGCTGAA	GGCAATTGATTCTGGAACCT
SSR-77[f]	TCTCCTCTTCTTGCTGAGCC	AGATTCACCCTTCAAGTTTCCTC
SSR-81[f]	AGTAATGAACCTGCCGCCTCTTT	TTGTCATTCTTGTGTTTTCCATCC
SSR-82[f]	CAAAATGAAGGAGAAAAGTGGACA	TGGCTTCATCTCAACCTTCCTTC
SSR-84[f]	AAGTAGATTGGTGAAAGGGAAGC	TCCTTCATTTTCTCCTCTTGGTT
SSR-87[f]	ATTCGACGACTCCAAAGCATA	CCTTGCTGGCCCTTCCTT
SSR-93[f]	ATACAGCAATTTTGAGAGGAGGAA	TTCTGTGCCTTCCCGAGTCA
SSR-100[f]	ACCCTTTACTACTTATTTACTCTC	ACATCCCCTTGCCATTTCTTC
SSR-102[b]	CTCAGAGCTGCGGTGGTGTCA	CCGGACGATCTTTCTTTCTTTCA
SSR-106[b]	CCCTCCCTCTTTCTCCTCTC	TCTGGGTTTTCTGTGTTCTCG
SSR-114[g]	TAACAGAAGCACCAAAACC	TCTAAACCCACCTCACAAC
SSR-119[d]	TTGCCATCATCGTTCATTCT	GCATAGTGTCGGTTGTGTTGTT
SSR-121[d]	CGACACTTTCTTTGGCACTC	AGACACCCACCCATCCAC
SSR-122[d]	CGTCTCGTTTCACGCTCTCT	GATCTGCATGTACTGGTGCTTC
SSR-138[d]	GCACTTCCCTCTCAACCAAC	ACTAGGACAGCAAATAGCATACACC
SSR-151[h]	TGGTTCAAGGTAATGTGGAAA	GGCCGTGATAATAAACAGCTA

[a] Rovelli et al. (2000)
[b] Combes et al. (2000)
[c] Coulibaly et al. (2003)
[d] Poncet et al. (2004)
[e] Baruah et al. (2003)
[f] Moncada et al. (2004)
[g] Leroy et al. (2005)
[h] Bhat et al. (2005)

The *C. canephora* species is divided into two varietal groups according to their diversity center and adaptive characteristics. The first group is called Conilon and is composed of genotypes that result in smaller leaves and fruits, less vigor, but greater tolerance to drought. The second group, called Robusta, is composed of higher and more vigorous genotypes with larger leaves and fruits, but is sensitive to drought (Ferrão 2013). Despite the significant difference that separates the two varietal groups, the classification of accessions in the germplasm collections is

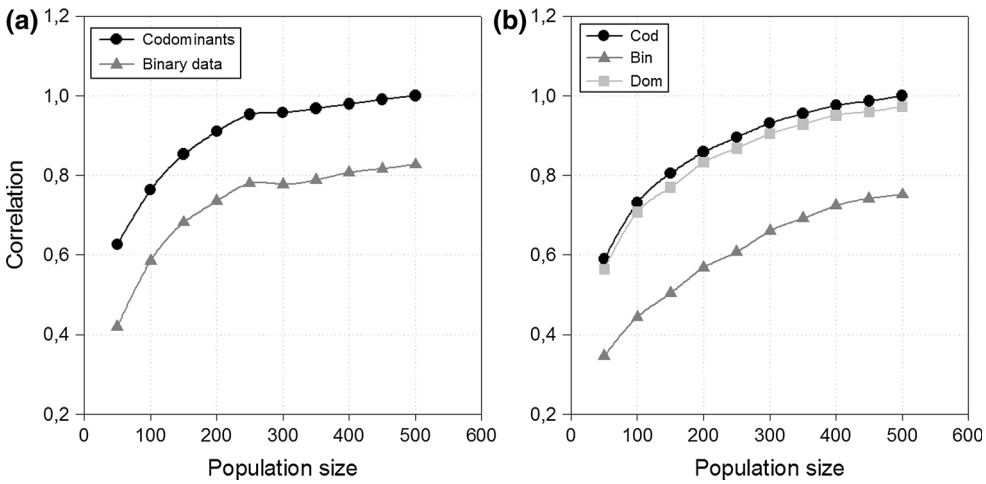


Fig. 1 **a** Scenario 1. SNP simulations using the codominant and binary data. **b** Scenario 2. SSR simulations using the codominant (*Cod*), binary (*Bin*) and most frequent allele format (*Dom*)

not an easy task. The natural form of reproduction (outcrossing) results in populations with high phenotypic amplitude and heterozygosity. In addition, it is necessary to consider the existence of natural crosses between the two varietal groups, which can complicate the discrimination of genotypes. Therefore, the correct evaluation of genetic diversity is important in a breeding program because it provides reliable information that can be used in the selection of promising genotypes.

In the diversity analysis, we observed that all methods were efficient at separating the groups of Conilon and Robusta. However, the efficiency of this classification ranged for each encoding approach (Fig. 2). These differences will be further discussed.

For the *Cod* and *Bin* encodings, similar results were observed (Fig. 2a, b, respectively). The dendrograms show similar structures, and the accessions were assigned to the Conilon, Robusta or natural hybrid groups. The main difference between these results was restricted within the groups, especially in the Conilon groups represented in Fig. 2 by the blue clade.

In the *Cod* evaluations, we observed three subgroups (Fig. 2a) in the Conilon group. Two of these subgroups were denominated as RO, and the accessions were collected in the State of Rondônia, Brazil. The third subgroup, called ES, was formed by accessions collected in the State of Espírito Santo, Brazil. The dendrograms show that most of the genotypes could be grouped according to their geographic origin. For the locality of Rondônia, we observed an additional structural organization in the two subgroups, which were designated RO.1 and RO.2. According to Souza (2011), the Rondônia germplasm was formed in the last four decades upon the introduction of seeds and clones from the States of Sao Paulo and Espírito Santo. Due to the greater similarity between the RO.2 and ES groups, it is assumed that the genotypes of these subgroups have the same origin (State of Espírito Santo). On the other hand, the RO.1 subgroup is formed by accessions that have originated in São Paulo, as they present their own characteristics that are quite distinct from those exhibited by the ES group.

The identification of three subgroups in the Conilon group was not obtained for the *Bin* encoding. The accessions maintained in this germplasm collection represent the genetic material cultivated and preserved from different Brazilian research institutions (Ferrão 2013; Souza 2011). Thus, it is expected that the genotypes from the same locality have similar molecular profiles because they share adaptive traits. This tendency was best seen in the *Cod* evaluation, where the majority of the ES accessions formed an isolated subgroup.

These results suggest that when we consider genotypes that are widely divergent, such as Robusta and Conilon accessions, the *Cod* and *Bin* evaluations are effective in varietal discrimination. However, in the case of similar accessions that belong to the same varietal group, such as the ES accessions, the *Cod* evaluations provide more detailed results and allow the discrimination of the accessions that share similar adaptive traits. The *Cod* advantage is the result of the evaluation method of codominant markers, where the similarity coefficients take into account the number of alleles shared and the number of loci studied. Another characteristic of the *Cod* evaluations is the possibility to work directly with the allele frequencies (Karp et al. 1997), which allows inferences about the genetic structure level, using Wright's (1965, 1978) *F* statistics and Nei's (1973) *G* statistics.

The distinction between the alleles in the dominant markers analysis can only be drawn if some assumptions on the data set are made, e.g., the existence of Hardy–Weinberg equilibrium (HWE) and linkage equilibrium (Bonin et al. 2007) in the population. However, when the germplasm accessions are analyzed, frequently, we cannot consider them as a population. Therefore, HWE cannot be assumed, which complicates the statistical inference about the allele frequency.

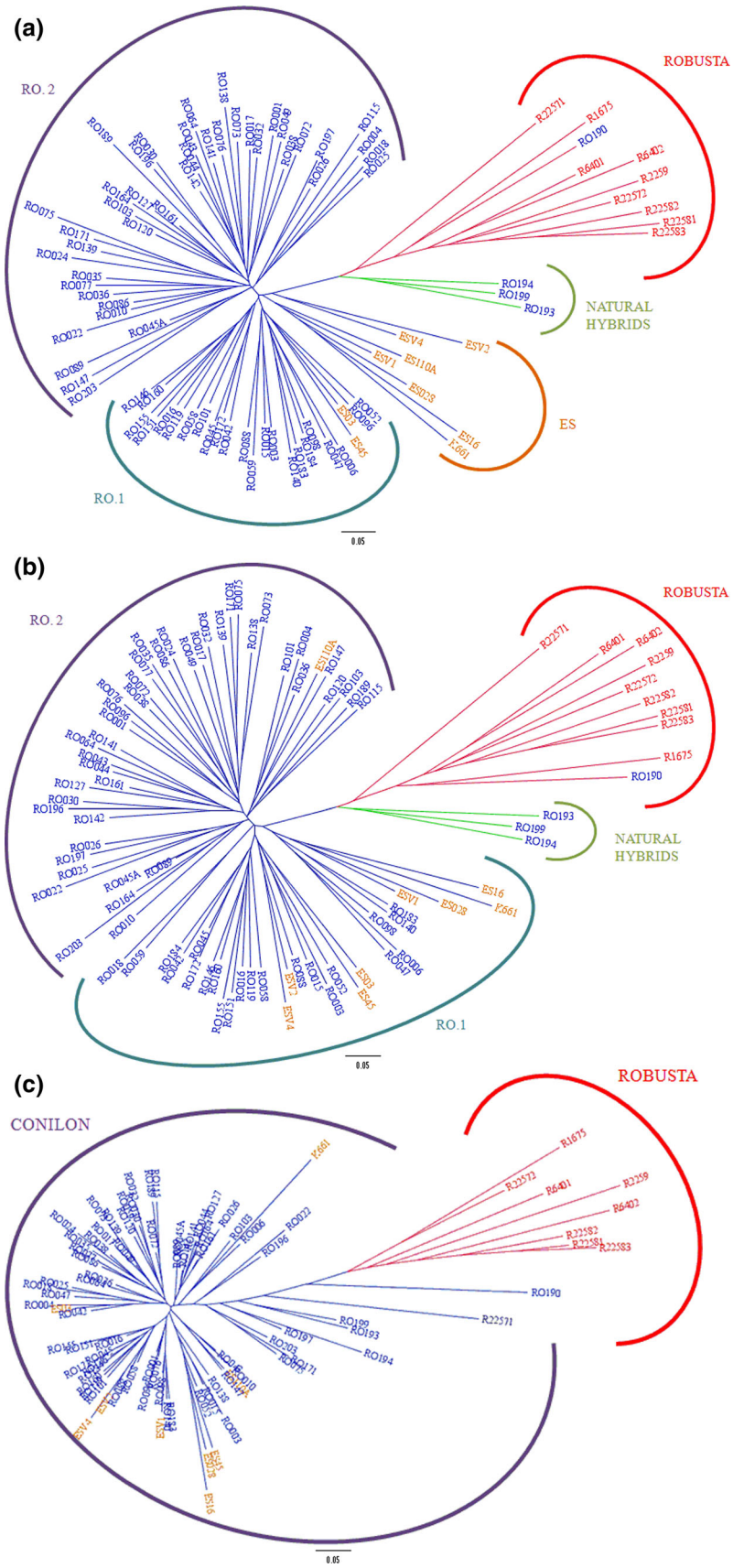


Fig. 2 Neighbor joining (NJ) dendrograms of the *Coffea canephora* accessions maintained in the Germplasm Bank of Embrapa Rondônia. *Blue*, *red* and *green* clades represent the following varietal groups: Conilon, Robusta and natural hybrids, respectively. The accession codes in orange are from the same locality and deserve special attention. Three encodings were used **a** *Cod*, **b** *Bin,* **c** *Dom*

Even though the *Cod* and *Bin* evaluations showed similar dendrograms, the *Dom* evaluation (Fig. 2c) did not provide robust results. In this analysis, the correct separation of the natural hybrids was not possible. Furthermore, the structuring of the Conilon and Robusta groups was different from the results obtained from the *Cod* analysis. The Robusta group was formed with fewer accessions, and the ES genotypes in the Conilon group did not show any grouping with adaptive logic. Moreover, in the *Dom* analysis, null values of the dissimilarity were observed, which prevented the discrimination of some accessions. Thus, despite the ease of procedure, this approach was inefficient and is not recommended for genetic diversity studies. In plant breeding, the use of this methodology can result in the loss of genetic gain, especially in programs that aim to identify heterotic groups and contrasting parents for the exploration of hybrid vigor.

The correct evaluation of diversity is a key factor in the selection of promising parents. If both parents are genetically similar, they share many genes or alleles in common. Thus, there is the expectation that divergent parents provide good hybrids according to the heterosis theories. In the management of genetic resources, the genetic diversity analysis may indicate the existence of false duplicates that are stored in the germplasm collection, resulting in the erroneous discard of promising materials. Moreover, a correct evaluation of germplasm is a valuable tool for breeding programs, especially at the beginning of the program when the work strategies are defined. For these reasons, the characterization and evaluation of the accessions should be accurate to help curators and researchers.

The results using the Bayesian approach are similar to those already presented. $K = 2$ was the highest value of ΔK for all evaluation modes, indicating that the accessions can be separated into two groups (Fig. 3). These results indicate that the highest hierarchy level of the accessions is related to the varietal group. Therefore, in the three evaluation methods, the Conilon and Robusta accessions could be distinguished from each other. Nevertheless, as in the cluster analysis, it was possible to separate the natural hybrids accurately only in the *Cod* (Fig. 3a) and *Bin* (Fig. 3b) analysis.

Detailed information about the subgroups inside the Conilon clade was obtained using $K = 4$, and the results were similar to those observed in the cluster analysis (Fig. 4). Using this K value in *Cod* (Fig. 4a), we observed a similarity between the genotype molecular profiles from the same locality and an efficient discrimination among the ES, RO.1, RO.2 and Robusta (ROB) subgroups. This observation confirms the hypothesis that the ES and RO.1 subgroups share adaptive traits. However, in the *Bin* evaluations (Fig. 4b), the ES and RO.1 subgroups were grouped together, assuming that this encoding mode was not efficient in the discrimination by locality.

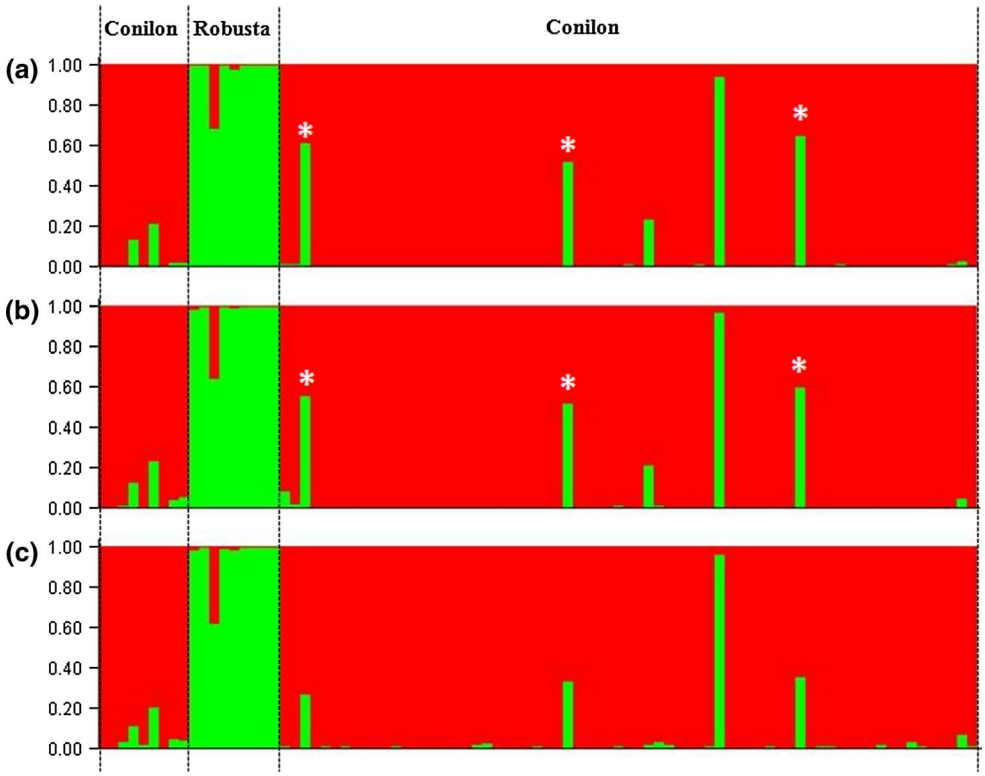


Fig. 3 Bar chart of the results from the STRUCTURE program used in the genetic diversity studies in *Coffea canephora*. The 87 genotypes are represented in the same order in Table 1 and are divided into two groups ($K = 2$) in accordance with the varietal groups **a** *Cod* evaluation, **b** *Bin* evaluation, and **c** *Dom* evaluation

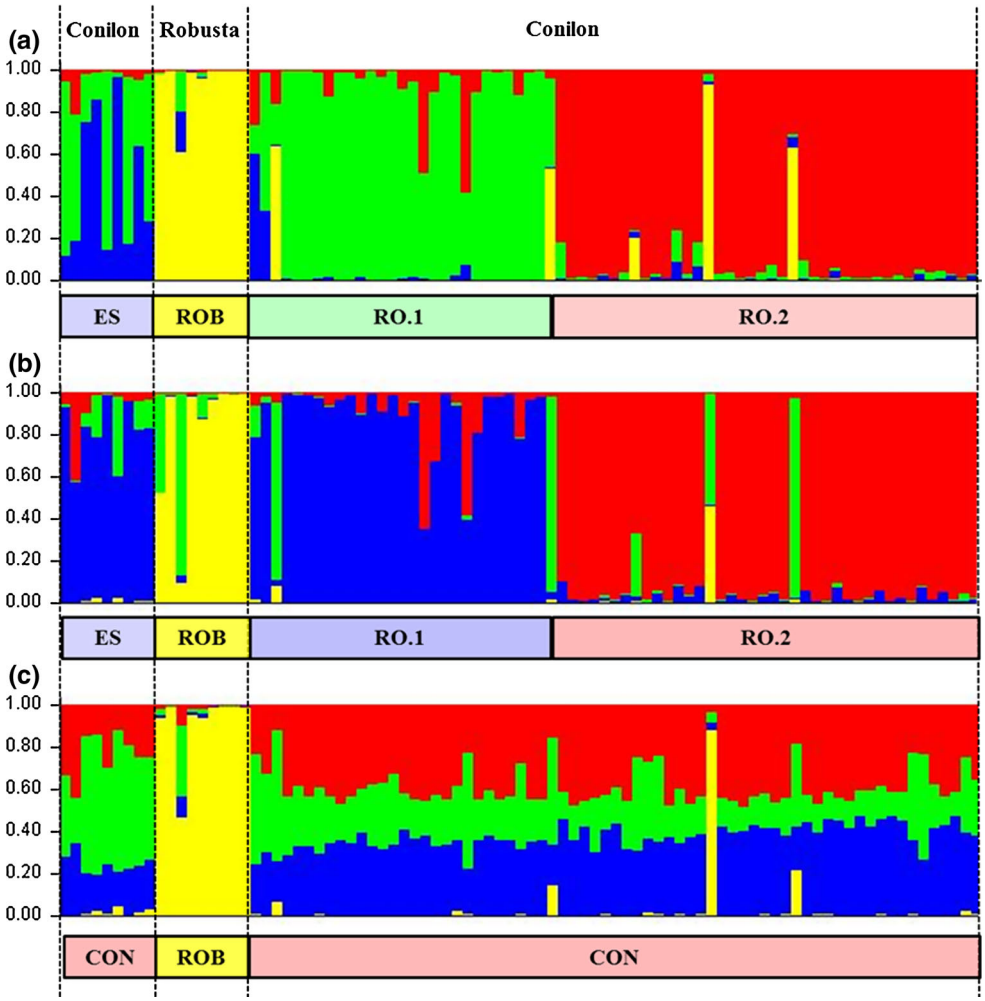


Fig. 4 *Bar chart* of the results from the STRUCTURE program used in the genetic diversity studies in *Coffea canephora*. The 87 genotypes are represented in the same order in Table 1 and divided into two groups in accordance with the varietal groups. $K = 4$ allowed the structuring of the accessions in the subgroups in accordance with the locality. The subgroups were designated as ES (accesses come from the State of Espírito Santo, Brazil), RO.1 and RO.2 (subdivisions of the genotypes that belong to the State of Rondônia, Brazil). The *color* of the subgroups indicates that there is a separation in the *Cod* (**a**), *Bin* (**b**) and *Dom* (**c**) evaluations. The clear differentiation of the three Conilon subgroups and the Robusta group (ROB) is only observed in the *Cod* evaluations

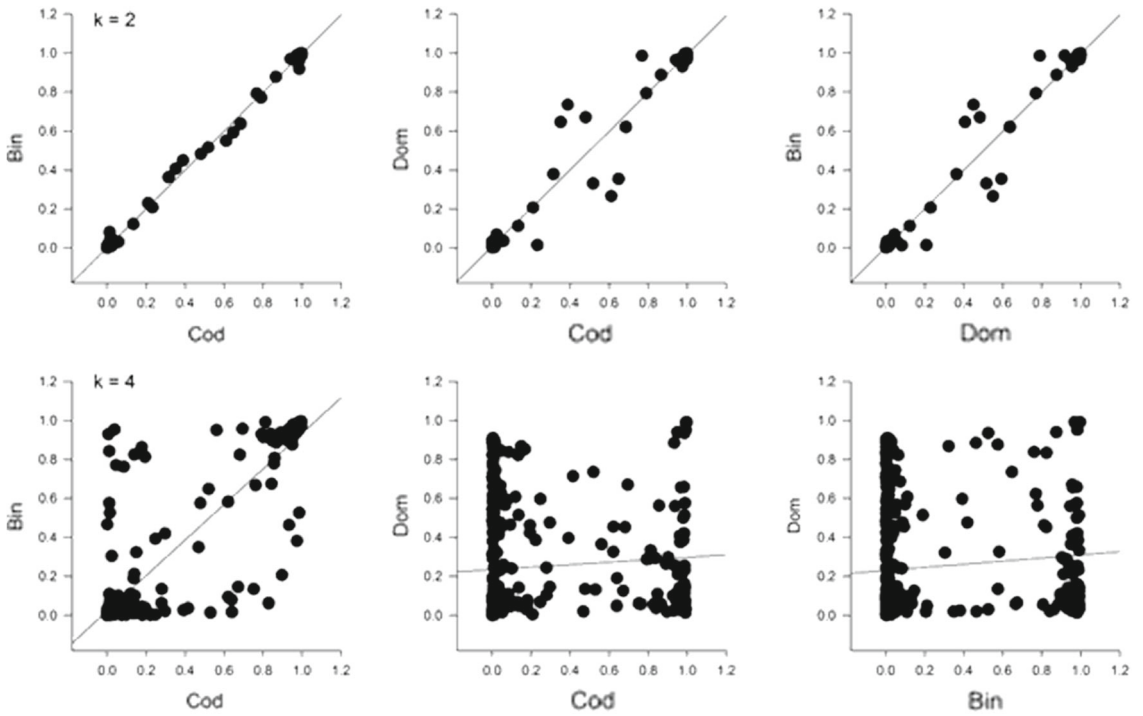
For the associative mapping studies, this information is critical because an inability to detect the population structures within the dataset results in the loss of the accuracy of associations and invalidates the statistical tests (Ewens and Spielman 1995). Finally, the *Dom* evaluations only allowed the separation between the Robusta and Conilon groups without providing any additional information about the possible subgroups. In addition, these evaluations showed identical molecular profiles among the accessions tested, which is a false indication of the duplicates in the germplasm collection.

The similarity of the results obtained from the Bayesian approach for $K = 2$ and $K = 4$ was quantified for the different evaluation modes. We measured the correlation of the ancestry coefficient values (Q) obtained between *Cod*, *Bin* and *Dom*. In the simple structuring ($K = 2$), the three evaluation modes were highly correlated. However, when using a higher and more detailed structuring level ($K = 4$), the correlation values decreased (Fig. 5).

The basic STRUCTURE algorithm was developed a priori for the multi-locus genotype analysis assuming linkage equilibrium and HWE within the populations. Subsequently, Falush et al. (2007) extended the original method using dominant markers and polyploid species and the binary scores. The high correlation values observed for $K = 2$ indicated that the information accessed by the encoding methods is similar. These results suggest that the extension of the original method was efficient. However, when we increased the informational requirements of the analyses using higher K values, we observed a decrease in the correlation values between the evaluation methods. Falush et al. (2007) reported a decrease in accuracy in dominant marker evaluations. According to these authors, this decrease is caused by genotypic ambiguity generated by the existence of recessive alleles.

Combined analysis of the dominant and codominant markers

The results discussed above confirm the loss of information and consequences in genetic resource management. The use of real data demonstrated that specific methodologies



Encodings	*Cod*	*Bin*	*Dom*
Cod	-	0.9006**	0.0831
Bin	0.9996**	-	0.1067
Dom	0.9925**	0.9944**	-

* Significant by *t* test (P=0.05) ; Significant by *t* test (P=0.01)

Fig. 5 Pearson correlation between the ancestry coefficient (Q) for the different encoding methods, using $K = 2$ (*below* the diagonal) and $K = 4$ (*above* the diagonal)

for codominant markers (*Cod*) are required. However, with the advancement of molecular biology, more than one type of markers is used for diversity studies to make the evaluations more accurate and reliable (Belaj et al. 2003; Gallego et al. 2005; Lamia et al. 2010). An important point in this study is that each marker type has its own characteristics that must be considered in the combined analysis. One practical example is the AFLP and SSR markers. The AFLP markers stand out to allow the analysis of a large number of loci per assay, and the SSR markers exhibit high informativeness and reproducibility *per* dataset. Thus, as the data are jointly analyzed, it is important to consider all these characteristics.

To consider the intrinsic characteristics of each marker, we proposed a methodology to study genetic diversity using with different types of markers. Each marker type was individually analyzed using the most appropriate genetic coefficient, e.g., in the dominant markers analysis, Jaccard (1908), Dice (1945) and simple-matching (Sokal and Michener, 1958) coefficients are the most commonly used. For the codominant markers, we highlighted the unweighted Index (Cruz et al. 2011), the Smouse and Peakall (1999) and the Kosman and Leonard coefficients (2005). Thus, for each marker type there will be one (dis)similarity matrix. Subsequently, these matrices are multiplied by a weighting index given by the following equation: $WI = L/N.\ i.QND$, where WI is the matrix weighting index, L is the number of loci accessed for each marker individually, N is the total number of loci accessed by all markers evaluated, i is the informativeness constant, and QND is the qualitative nature of the data. The WI approach adds important information to the analysis, e.g., genome coverage, informativeness and data quality. Thus, the most complete molecular markers will be weighted with the highest scores and have greater representativeness in the genetic diversity studies.

In the WI calculation, it is difficult to define the constant values of informativeness (i) and qualitative nature of the data (QND). Considering the particularities of each marker, we propose i values for dominant and codominant markers, and within the codominant markers, for multi-allelic or biallelic (Table 3). Multi-allelic and codominant markers

Table 3 Informativeness values (*i*) with different types of molecular markers

Class	Evaluation	Molecular markers	Informativeness values (*i*)
Codominant	Multi-allelic	SSR	1.00
	Biallelic	SNP and DaRTs	0.75
Dominant	Binary	AFLP, RAPD and ISSR	0.60

Table 4 Values scale of quality of marker (QM)

Scale	Quality of markers
1.00	Good quality marker—single and strong band/peak
0.75	Faint band or lower peak
0.50	Marker/band with stuttering
0.25	Difficult to score (needs special efforts to visualize)

Varshney et al. (2007)

Table 5 Documentation capability (DC) and the percent of reproducibility of the fragment(s)/band(s)/peak(s) (PR) values for the different types of markers

Parameter	Molecular markers			
	SNP and DaRTs	SSR	AFLP and ISSR	RAPD
DC	1.00	0.75	0.50	0.25
PR	1.00	1.00	0.50	0.25

Adapted Varshney et al. (2007)

were weighted with higher grades because they are more informative than the other types of markers (Ferrão 2013; Gallego et al. 2005; Lamia et al. 2010; Poncet et al. 2004; Russell et al. 1997).

The concept of QND was presented by Varshney et al. (2007) and is calculated with the following formula: QND = DC × QM × PR, where DC is the documentation capability, QM is the quality of the marker and PR is the Percent of Reproducibility of the fragment(s)/band(s)/peak(s) of the given marker system across the laboratories. The QM values range over the primer combinations for each marker type. For the QND final value, the QM average value is used. Table 4 presents the range of values for this parameter.

The DC and PR values are presented in Table 5 and were suggested taking the characteristics of each maker into consideration. For the DC parameter, lower values were used for the markers that analyze multiple loci per assay, such as AFLP, due to the large number of bands/peaks. This feature makes the interpretation in the automated genotyping systems difficult. On the other hand, molecular markers based on hybridization methods using DNA solid platforms obtained higher values. This is the case for SNP and DArT, which are automatically documented in a "digital fashion" that is convenient for storage in the database (Jaccoud et al. 2001; Varshney et al. 2007). Specific loci markers, such as SSR, are easily evaluated because they have at most two bands/peaks per locus in the diploid organism. However, we suggest an intermediate DC value for them because some artifacts that occur in the assay interfere with the analysis (Guichoux et al. 2011).

The PR value represents the reliability of each technique. In the germplasm characterization, this is a valuable parameter because it indicates the reproducibility of the results and can be shared among different laboratories. In this way, the SNP and DArT are considered the most robust markers, while the RAPD markers are considered the least reliable. We believe that for genetic analyses with repetitions and rigorous data controls, it is possible to disregard the QND parameter in the PI calculation.

The weighting index approach allows for the simultaneous use of more than one class of molecular markers in genetic diversity. Beyond adding informativeness in the analysis, another advantage of this method is the obtainment of a single outcome that summarizes all of the information into a single (dis) similarity matrix and dendrogram. On the other hand, traditional methods usually involve an individual analysis for each molecular marker, resulting in several dendrograms. In this context, it is difficult to draw conclusions about the divergence because different molecular markers access different levels of information in the genome and cannot lead to similar results. In this sense, the use of the weighting index makes the results clearer and facilitates the conclusions about the study.

For polyploid organisms, the genetic diversity studies are characterized by complex patterns of inheritance and difficulty or impossibility in determining the exact number of copies of each allele (Serang et al. 2012). Some solutions have been proposed to minimize this problem. The most widely used solution for codominant markers is the encoding of data using binary form evaluations. The consequences of this adaptation in genetic diversity studies have not been reported thoroughly. In this paper, we showed that the encodings in diploid organisms must be avoided because they result in the loss of information and lower accuracy. If we extrapolate this observation to polyploid studies, we can conclude that the problem persists and encoding data are a problem. The major difference between these two cases is that encoding in polyploids is necessary because it is not possible to determine the exact number of copies of each allele. Thus, it is necessary to develop robust methodologies that could solve this problem.

Accordingly, some solutions have been proposed with an emphasis on a study performed by Serang et al. (2012).

This group presented a Bayesian graphic model for SNP genotyping in which the genotypes can be inferred in populations where the ploidy level is unknown. These concepts have been implemented in the SuperMASSA® software and are an excellent alternative for polyploid studies with SNPs. Methodologies and specific polyploidy software should also be used in genetic analyses. For the diversity studies, we highlight the FDASH (Obbard et al. 2006), TETRASAT (Markwith et al. 2006) and ATETRA (VAN Puvvelde et al. 2010) programs.

Finally, the approach presented in this article may be expanded in a more general context, which involves the use of next generation DNA sequencing (NGS). In this scenario, modern methods of genotyping, such as GBS (Elshire et al. 2011) and RAD-seq (Baird et al. 2008), can be used with traditional markers (AFLP and RAPD SSR, for example). The advantage of the weighting index is due to the SNP data from these modern platforms being weighted with higher scores than traditional data. This occurs in two components of expression. The first component is related to the number of sampled loci (*L*). Studies with GBS, for example, are able to generate hundreds of SNPs depending of the germplasm that was evaluated (Polland and Rife 2012). These numbers are much higher than the results from studies with AFLP and SSR, which are able to sample hundreds of loci. The second component is associated with the qualitative nature of the data (QND). In modern approaches, the reproducibility (PR), data quality (QM) and documentation capabilities (DC) of the SNPs are better than in traditional methods, mainly because all the steps are automated using accurate methodologies for next generation sequencing.

Conclusions

1. For simulated data, the encoding methods resulted in a loss of information for the two proposed scenarios. This is a problem for studies involving multi-allelic and biallelic markers, suggesting that the encoding data must not be used in genetic diversity studies.
2. In studies involving real data using SSR, the encoding data were inefficient in genetic breeding studies, especially those that aim at the identification of heterotic groups and evaluation of genetic resources stored in germplasm collections. As to the *C. canephora* studies, the encoding data were not effective in discriminating all the subgroups by any of the approaches used (cluster analysis and Bayesian approach).
3. Compared to the genetic diversity studies using different types of markers, when performing the joining of data, it is important that the intrinsic features of each assay be considered. Thus, it is important to compute the informativeness, coverage and quality of the markers. The weighted index proposed in this paper is a methodology that takes into account all these factors, making it an important tool in genetic studies.

Acknowledgments The authors thank Dr. Romário G. Ferrão, Dr. Abrãao C. Verdin-Filho and Paulo Volpi for giving us additional coffee samples from the Capixaba Research Institute—Technical Assistance and Rural Extension (Instituto Capixaba de Pesquisa, Assistência Técnica e Extensão Rural -Incaper). We also thank Milton M. Santos, João Maria Diocleciano and Gilvan O. Ferro for the technical support at Embrapa Experimental Station, in Rondônia, and, Rejane L. Freitas, Telma Fallieri and Tesfahun A. Sotetaw for the technical support at UFV laboratory, in Viçosa. This work was financially supported by Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café, Agrofuturo—Embrapa and National Council of Scientific and Technological Development (CNPq).

References

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3:e3376. doi:10.1371/journal.pone.0003376

Baruah A, Naik V, Hendre PS, Rajkumar R, Rajendrakumar P, Aggarwal RK (2003) Isolation and characterization of nine microsatellite markers from *Coffea Arabica* L., showing wide cross-species amplifications. Mol Ecol Notes 3:647–650

Belaj A, Satovic Z, Cipriani G, Baldoni L, Testolin R, Rallo L, Trujillo I (2003) Comparative study of the discriminating capacity of RAPD, AFLP and SSR markers and of their effectiveness in establishing genetic relationships in olive. Theor Appl Genetics 107:736–744

Bhat PR, Krishnakumar V, Hendre PS, Rajendrakumar P, Varshney RK, Aggarwal RK (2005) Identification and characterization of expressed sequence tags-derived simple sequence repeats markers from robusta coffee variety 'CxR' (an interspecific hybrid of *Coffea canephora x Coffea congensis*). Mol Ecol Notes 5:80–83

Bonin A, Ehrich D, Manel S (2007) Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists. Mol Ecol 16:3737–3758

Bruvo R, Michiels NK, D'Souza TG, Schulenburg H (2004) A simple method for the calculation of microsatellite genotype distances irrespective of ploidy level. Mol Ecol 13:2101–2106

Combes MC, Andrzejewski S, Anthony F, Bertrand B, Rovell P, Graziosi G, Sashermes P (2000) Characterization of microsatellites loci in *Coffea arabica* and related coffee species. Mol Ecol 9:1171–1193

Coulibaly I, Revol B, Noirot M, Poncet V, Lorieux M, Carasco-Lacombe C, Minier J, Dufour M, Hamon P (2003) AFLP and SSR polymorphism in a *Coffea* interspecific backcross progeny [(*C heterocalyx x C. canephora) x C. canephora*]. Theor Appl Genet 107:1148–1155

Cruz CD (2013) GENES—a software package for analysis in experimental statistics and quantitative genetics. Acta Sci 35:271–276

Cruz CD, Medeiros FF, Pessoni LA (2011) Biometria aplicada ao estudo de diversidade genética. Viçosa, MG

De Silva HN, Hall AJ, Rikkerink E, McNeilage MA, Fraser LG (2005) Estimation of allele frequencies in polyploids under certain patterns of inheritance. Heredity 95:327–334

Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26:297–302
Diniz EC, Sakiyama NS, Lashermes P, Caixeta ET, Oliveira ACB, Zambolim EM, Loureiro ME, Pereira AA, Zambolim L (2005) Analysis of AFLP markers associated to the Mex-1 resistance locus in Icatu progenies. Crop Breed Appl Biotechnol 5:387–393
Earl D, vonHoldt B (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genetics Resour 4:359–361
Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-bysequencing (GBS) approach for high diversity species. PLoS One 6:e19379. doi:10.1371/journal.pone.0019379
Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–2620
Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. Mol Ecol Notes 7:574–578
Ferrão LFV, Caixeta ET, Souza FD, Zambolim EM, Cruz CD, Zambolim L, Sakiyama NS (2013) Comparative study of different molecular markers for classifying and establishing genetic relationships in *Coffea canephora*. Plant Syst Evol 299:225–238
Gallego FJ, Perez MA, Nunez Y, Hidalgo P (2005) Comparison of RAPDs, AFLPs and SSR markers for the genetic analysis of yeast strains of *Saccharomyces cerevisiae*. Food Microbiol 22:561–568
Ramos HCC, Pereira MG, Goncalves LSA, do Amaral AT, Scapim CA (2011) Comparison of multiallelic distances for the quantification of genetic diversity in the papaya. Acta Sci Agron 33:59–66
Jaccard P (1908) Nouvelles recherches sur la distribution florale. Bull Soc Vaudoise Des Sci Nat, 223–270
Jaccoud D, Peng K, Feinstein D, Kilian A (2001) Diversity arrays: a solid state technology for sequence information independent genotyping. Nucl Acids Res 29:E25
Karp A, Kresovich S, Bhat K, Ayad W, Hodgkin T (1997) Molecular tools in plant genetic resources conservation: a guide to the technologie. International Plant Genetic Resources Institute, Rome
Kosman E, Leonard KJ (2005) Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. Mol Ecol 14:415–424
Lamia K, Hedia B, Jean-Marc A, Neila TF (2010) Comparative analysis of genetic diversity in Tunisian apricot germplasm using AFLP and SSR markers. Sci Horticult 127:54–63
Laurentin H (2009) Data analysis for molecular characterization of plant genetic resources. Genetic Resour Crop Evol 156:277–292
Leroy T, Marraccini P, Dufour M, Montagnon C, Lashermes P, Sabau X, Ferreira LP, Jourdan I, Pot D, Andrade AC, Glaszmann JC, Vieira LG, Piffanelli P (2005) Construction and characterization of a *Coffea canephora* BAC library to study the organization of sucrose biosynthesis genes. Theor Appl Genet 111:1032–1041
Markwith SH, Stewart DJ, Dyer JL (2006) TETRASAT: a program for the population analysis of allotetraploid microsatellite data. Mol Ecol Notes 6:586–589
Missio RF, Caixeta ET, Zambolim EM, Zambolim L, Cruz CD, Sakiyama NS (2010) Polymorphic information content of SSR markers for *Coffea* spp. Crop Breed Appl Biotechnol 10:89–94
Mohammadi SA, Prasanna BM (2003) Analysis of genetic diversity in crop plants—salient statistical tools and considerations. Crop Sci, 1235–1248
Moncada P, McCouch S (2004) Simple sequence repeat diversity in diploid and tetraploid *Coffea* species. Genome 47:501–509
Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70:3321–3323
Obbard DJ, Harris SA, Pannell JR (2006) Simple allelic-phenotype diversity and differentiation statistics for allopolyploids. Heredity 97:296–303
Polland JA, Rife T (2012) Genotyping-by-sequencing for plant breeding and genetics. The Plant Genome 5:3
Poncet V, Hamon P, Minier J, Carasco C, Hamon S, Noirot M (2004) SSR cross-amplification and variation within coffee trees (*Coffea* spp.). Genome 47:1071–1081
Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959
Rambaut A (2006) Tree figure drawing tool version 1.3.1. Institute of Evolutionary Biology, University of Edinburgh, UK
Guichoux E, Lagache L, Wagner S, Chaumeil P, Leger P, Lepais O, Lepoittevin C, Malausa T, Revardel E, Salin F, Petit RJ (2011) Current trends in microsatellite genotyping. Mol Ecol Resour 11:591–611
Rovelli P, Mettulio R, Anthony F, Anzuetto F, Lashermes P, Graziosi G (2000) Microsatellites in *Coffea arabica* L. In: Sera T, Soccol CR, Pandey A and Roussos S (eds) Coffee biotechnology and quality. Kluwer Academic Publishers, Dordrecht, pp 123–133
Russell JR, Fuller JD, Macaulay M, Hatz BG, Jahoor A, Powell W, Waugh R (1997) Direct comparison of levels of genetic variation among barley accessions detected by RFLPs, AFLPs, SSRs and RAPDs. Theor Appl Genet 95:714–722
Serang O, Mollinari M, Garcia AA (2012) Efficient exact maximum a posteriori computation for bayesian SNP genotyping in polyploids. PLoS One 7:e30906
Smouse PE, Peakall R (1999) Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. Heredity 82:561–573
Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. Univ Kansas Sci Bull, 1409–1438
Souframanien J, Gopalakrishna T (2004) A comparative analysis of genetic diversity in blackgram genotypes using RAPD and ISSR markers. Theor Appl Genet 109:1687–1693
Souza FF (2011) Estudos sobre a diversidade, estrutura populacional, desequilíbrio de ligação e mapeamento associativo em *Coffea canephora* Pierre ex Froehner, Dissertation of Universidade Federal de Viçosa, Viçosa, Brazil
Ewens WJ, Spielman, RS (1995) The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet 57:455–64
VAN Puvvelde K, VAN Geert A, Triest L (2010) atetra, a new software program to analyse tetraploid microsatellite data: comparison with tetra and tetrasat. Mol Ecol Resour 10:331–334
Varshney RK, Chabane K, Hendre PS, Aggarwal RK, Graner A (2007) Comparative assessment of EST-SSR, EST-SNP and AFLP markers for evaluation of genetic diversity and conservation of genetic resources using wild, cultivated and elite barleys. Plant Sci 173:638–649
Wright S (1965) The interpretation of population-structure by *F*-statistics with special regard to systems of mating. Evolution, 395–420
Wright S (1978) Evolution and the genetics of populations. Univ. Chicago Press, Chicago