Revista Ciência Agronômica, v. 46, n. 4, p. 836-845, out-dez, 2015
Centro de Ciências Agrárias - Universidade Federal do Ceará, Fortaleza, CE
www.ccarevista.ufc.br

Artigo Científico
ISSN 1806-6690

Establishing a core collection from the integration of morpho-agronomical, phytopathological and molecular data[1]

Estabelecimento de coleção nuclear a partir da integração de dados morfoagronômicos, fitopatológicos e moleculares

Francielle Alline Martins[2]*, Derly José Henriques da Silva[3] e Pedro Crescêncio Souza Carneiro[4]

ABSTRACT - The aim of this study was to establish and compare, as to their representativeness, core collections obtained from quantitative data, multicategoric, molecular and collections that covering all this information simultaneously. Ten subcollections were established from 67 tomato accessions of the Germplasm Bank of the Universidade Federal de Viçosa (BGH-UFV), characterized according to 19 quantitative traits, 30 multicategoric characters, 52 ISSR loci and to the reaction to three pathogens. These subcollections were defined by the combination of the nature of data collected and the sampling rate. The COD-20 subcollection stood out in 20% intensity of sampling by has higher rates of coincidence amplitude followed by more appropriate values of variance. At 30% intensity, subcollection MOL-30 was as efficient as the subcollection COD-30 when considering only the rates of coincidence of the amplitude and the variances. However, the graphical analysis of the variability showed a slight superiority of subcollection COD-30 in maintaining the variability, especially regarding multicategoric characters. So whenever data from different sources are available, should be prioritized the establishment of core collections from the integration of these data, since these were more representative when the amplitude coefficient, variance, and retention index of variability, are regarded simultaneously.

Key words: Retention rate variability. Amplitude coefficient. Germplasm. Variance.

RESUMO - O objetivo deste estudo foi estabelecer e comparar, quanto à representatividade, coleções nucleares obtidas a partir de dados quantitativos, multicategóricos, moleculares e coleções que contemplem todas essas informações, simultaneamente. Foram estabelecidas 10 subcoleções a partir de 67 acessos de tomateiro do Banco de Germoplasma de Hortaliças da Universidade Federal de Viçosa (BGH-UFV), caracterizados quanto a 19 caracteres quantitativos, 30 multicategóricos, 52 locos ISSR e reação a três patógenos. Essas subcoleções foram definidas pela combinação entre a natureza dos dados avaliados e a intensidade de amostragem. A subcoleção COD-20 destacou-se a 20% de intensidade de amostragem por possuir maiores índices de coincidência da amplitude acompanhados de valores de variância mais adequados. A 30% de intensidade, a subcoleção MOL-30 foi tão eficiente quanto à subcoleção COD-30, quando se considerou apenas os índices de coincidência da amplitude e as variâncias. Entretanto, a análise gráfica da variabilidade mostrou uma ligeira superioridade da subcoleção COD-30 em manter a variabilidade, principalmente em relação aos caracteres multicategóricos. Assim, sempre que dados de diferentes naturezas estiverem disponíveis, deve-se priorizar o estabelecimento de coleções nucleares a partir da integração dos mesmos, uma vez que essas se mostraram mais representativas, quando considerados o coeficiente de amplitude, variância e o índice de retenção da variabilidade, simultaneamente.

Palavras-chave: Índice de retenção da variabilidade. Coeficiente de amplitude. Germoplasma. Variância.

DOI: 10.5935/1806-6690.20150072
* Autor para correspondência
[1]Recebido para publicação em 21/04/2014; aprovado em 28/08/2015
Parte da Tese de Doutorado do primeiro autor apresentada ao Programa de Pós-Graduação em Genética e Melhoramento da Universidade Federal de Viçosa
[2]Centro de Ciências da Natureza, Universidade Estadual do Piauí, Teresina-PI, Brasil, franufv@yahoo.com.br
[3]Departamento de Fitotecnia, Universidade Federal de Viçosa, Viçosa-MG, Brasil, derly@ufv.br
[4]Departamento de Biologia Geral, Universidade Federal de Viçosa, Viçosa-MG, Brasil, carneiro@ufv.br

INTRODUCTION

The emphasis given to the importance of preserving genetic resources has led to the formation and maintenance of large germplasm collections around the world. However, the large size of these collections has often been an obstacle to their use, conservation and management (VASCONCELOS *et al*., 2010). With the objective of enhancing utilisation and accessibility, and minimising maintenance difficulties, Frankel (1984) proposed the concept of the core collection, as being a subsample derived from a set of germplasm chosen to represent, with minimum redundancy, the maximum genetic variability of the initial collection or base of a particular species.

Establishing a core collection requires an integrated effort involving curators, breeders and geneticists, to define the size and the accessions that will make up the collection (ABADIE *et al*., 2005). Much work is done in order to establish core collections, and several strategies have been outlined for their formation, which include from simple random, non-random, and stratified sampling (UPADHYAYA *et al*.,2007; XU *et al*., 2006), to more sophisticated sampling methods (JANSEM; VAN HINTUM, 2007; VASCONCELOS *et al*., 2007; WANG *et al*., 2007).

Whatever the methodology adopted, it will be based on the information available concerning germplasm diversity, which can take different forms, such as passport data (genealogical, geographical origin) (DWIVEDIL; UPADHYAYA; HEDGE, 2005; LI *et al*., 2004), phenological data (UPADHYAYA *et al*., 2007), morpho-agronomical data (WANG *et al*., 2011; ZEWDIE; TONG; BOSLAND, 2004), or molecular data (HAO *et al*., 2006; WANG *et al*., 2006).

Studies of genetic diversity show that there is not always a correlation between the dissimilarity distances established from the different types of data, so that the diversity observed from one set of characteristics, cannot always be extrapolated to the rest (GOMES, 2007; MARTINS *et al*., 2011). It is therefore prudent to consider the maximum amount of information to characterise the germplasm when establishing a core collection.

In general, studies of genetic diversity have been carried out, and core collections established, based on quantitative, qualitative or molecular characteristics alone. Uniting all the data into a single study has been hampered by the absence of germplasm banks that would contain such a detailed characterisation of the germplasm, due to limited financial and human resources (MARTINS *et al*., 2011).

Although scarce, methodologies for the integration of different types of data into a single analysis are required to establish more effective core collections, especially when the primary purpose is the preservation of genetic variability.

The aim of this study therefore, was to establish and validate core collections obtained from quantitative, multi-category or molecular data, and from collections that include all of this information.

MATERIAL AND METHODS

The study was carried out on a set of data of 67 tomato accessions from the Vegetable Germplasm Bank of the Federal University of Viçosa (BGH-UFV); all had been previously characterised for 19 quantitative features, 30 multi-category features and 53 ISSR loci - determined by Aguilera *et al*. (2011) - and for their reaction to *Alternaria solani*, *Pseudomonas syringae* pv. Tomato, and to the begomovirus, Tomato yellow spot (ToYSV).

Quantitative characteristics were evaluated for different development phases of the tomato. At the seedling stage, the diameter of the hypocotyl and length of the cotyledon were evaluated. At the vegetative stage, the thickness of the main petiole and the lengths of the leaf and internode were evaluated. In the fruit, the following characteristics were assessed: fruit length, width of the fruit and central axis, thickness of the endocarp, number of locules, total soluble solids content, total titratable acidity and organoleptic quality. In addition, agronomic characteristics such as the weight and number of good fruit, the weight and total number of fruit, average fruit weight and precocity index were evaluated.

Data for reaction of the accessions to *A. solani* and *P. syringae* are by nature quantitative, as they are obtained respectively from measurement of the leaf area damaged by the fungus, and by counting the total number of bacterial pustules on each plant. The reaction of the accessions to ToYSV was characterised into five classes: highly resistant, resistant, moderately resistant, susceptible and highly susceptible. The data from this evaluation were analysed together with the multi-category data: hypocotyl colour; type of plant growth; density of pilosity on the stem and foliage; attitude and type of leaf; type of corolla; external colour of the immature fruit; the presence and frequency of green shoulder on the fruit; the shape, homogeneity and size of the fruit; colouration and colouration intensity in the mature fruit; secondary fruit shape and shape of the shoulder; size of the area of corking around the pedicel scar; ease of removal of the epicarp; colour of the epicarp; colour and colour intensity of the mesocarp; shape of the cross section of the fruit and stylar scar; shape of the distal extremity of

the fruit; condition of the stylar scar; colour of the central axis; and radial and concentric cracking.

The accessions were previously separated or stratified by commercial group: Saladinha, Santa Cruz, Italian or Saladette, Apple or Persimmon and Cherry, according to an analysis of the shape of the fruit, assessed by grading (IPGRI, 1996) (Table 1).

Ten sub-collections of tomato accessions from BGH-UFV were evaluated, defined by the combination of the type of data evaluated and a sampling density of 20 or 30%, representing 14 and 20 accessions per collection respectively. A logarithmic strategy was employed for deciding the number of entries selected for each stratum, calculated from the expression:

$$NAS = \frac{\log a_i}{\sum_{i=1}^{ns} \log a_i} .nt \quad (1)$$

where: *NAC* is the number of accessions sampled per stratum; a_i is the number of accessions of the i-th class; *nt* is the total number of accessions sampled, defined by the sampling density; and *ns* is the number of strata.

The choice of accessions to comprise each of the sub-collections was based on diversity analysis within each stratum. Dissimilarity matrices between accessions within each stratum were therefore obtained from the quantitative data, either multi-category or molecular, whether integrated by matrix addition or by encoding the quantitative data into multi-category data through a strategy of equal division of the amplitude into three classes, as described by Martins *et al*. (2011). In obtaining the dissimilarity matrices, the standardised mean Euclidean distance was used for the quantitative data, and the arithmetic complements of the simple index of coincidence for the multi-category data, and of the Jaccard similarity coefficient for the molecular data.

For each dissimilarity matrix from each stratum, the maximum distance value was found and converted into a similarity value using the equation:

$$s = 100 - \left(\frac{100.d}{d_{max}}\right) \quad (2)$$

where: *s* is the similarity; *d* is the distance value between the i and i' individuals on the dissimilarity matrix; and d_{max} is the maximum distance value.

In this way, the maximum distance was equated to zero. Due to the dissimilarity matrix becoming a similarity matrix, when grouping accessions by the Tocher method,

Table 1 - Classification of 67 tomato accessions from the Vegetable Germplasm Bank of UFV (BGH-UFV) by commercial group

BGH-UFV	CG	BGH-UFV	CG	BGH-UFV	CG	BGH-UFV	CG
166	SC	990	SAL	1532	SAL	2211	SAL
181	SAL	991	SC	1538	SC	2213	SAL
279	SC	992	SC	1706	SC	2214	SAL
322	SAL	993	SC	1708	SC	2216	AP
349	SAL	994	SC	1985	SAL	2219	SAL
468	SAL	997	SAL	1987	SAL	2223	AP
489	SAL	1019	SAL	1988	SAL	2229	SAL
773	CHE	1020	SC	1989	SAL	2234	AP
850	SAL	1211	SAL	1990	ITA	3472	SAL
970	SC	1214	SAL	1991	SAL	4006	SC
975	SC	1254	CHE	1992	SAL	4035	SC
978	SAL	1282	SAL	1993	SAL	4053	SC
980	CHE	1485	SAL	2119	SAL	4054	ITA
981	SAL	1490	SC	2202	SAL	4055	SC
985	SAL	1497	SC	2203	SAL	4206	SC
987	SAL	1498	SC	2205	SAL	4309	SC
989	SAL	1499	SAL	2208	SAL		

BGH-UFV - Identification number of the accessions in the Vegetable Germplasm Bank of the Federal University of Viçosa; CG - Comercial group; SAL - Saladinha; SC - Santa Cruz; ITA - Italiano or Saladette; AP - Apple ou Persimmon; CHE - Cherry

an inverse result was produced, i.e. the most divergent genotypes formed clusters (VASCONCELOS *et al*., 2007). Selection of those accessions that made up the sub-collections was carried out according to the clustering sequence of the Tocher method, until the number of accessions reached the predetermined selection number for each stratum, according to sampling density (OLIVEIRA *et al*., 2010).

The process of validation of the sub-collections involved comparing them to the initial collection. Comparisons were made taking into consideration the amplitude coincidence index (AC), and the variance for each group of characteristics, whether quantitative, multi-category or molecular. In addition, a graphical analysis of the variability was proposed, for which the sub-collections were also compared using a retention of variability index (RVI).

The amplitude coincidence index (AC) for each sub-collection was obtained for each group of characteristics, whether quantitative, multi-category or molecular, by means of the equation (HU; ZHU; XU, 2000; WANG *et al*., 2007):

$$AC = \frac{1}{n}\sum_{i=1}^{n}\frac{A_iSC}{A_iCI} \quad (3)$$

where: AC is the amplitude coincidence index; A_iSC is the amplitude of the i-th characteristic in the sub-collection; *AiCI* is the amplitude of the i-th characteristic in the initial collection; and *n* is the number of characteristics for a particular group.

The variance was estimated for each characteristic, both in the initial collection and the core sub-collections. The average of these variances was obtained for each set of characteristics in each of the sub-collections and in the main collection. All comparisons between variances were carried out by F-test (SNEDECOR; COCHRAN, 1980).

To assess the representativeness of the core sub-collections as to retention of variability, the encoded quantitative data and the multi-category data were recoded into binary data, i.e. each class was regarded as one characteristic and the accessions were rated as 1 when belonging to that class, and 0 when not.

The frequency of accessions in each class was estimated for all the core sub-collections. The class of a characteristic, when present in all individuals of a sub-collection, was considered a fixed characteristic, while those classes with zero frequency in the sub-collections were considered as extinct characteristics. In short, this showed the fixation or loss of alleles for the different phenotypes of a characteristic. Once the frequency of each characteristic in the initial collection and in the sub-collections was estimated, the values were plotted on a graph and the retention of variability index (RVI) estimated:

VRI = (number of classes kept in the

sub-collections/total number of classes) x 100 (4)

RESULTS AND DISCUSSION

In the samples for a density of 20%, the number of accessions in the Saladinha, Santa Cruz, Italian, Apple and Cherry strata were 5, 4, 1, 2 and 2 respectively. While in the samples at 30% density, the number of accessions for the same strata were 8, 6, 2, 2 and 2 respectively.

From the grouping for each stratum by the inverse Tocher method based on the dissimilarity matrices obtained from the different types of data, and considering the adopted sampling densities of 20% and 30%, 10 core sub-collections were established (Table 2).

Only the 980, 2216 and 2234 accessions from the Cherry, Apple and Apple strata respectively, are present in all the sub-collections formed. For those strata with a higher number of accessions, Saladinha and Santa Cruz, no accessions were seen common to all the sub-collections. The differences found arise from the use of different types of data in establishing the sub-collections. This suggests that sub-collections based on isolated groups of characteristics do not cover the genetic diversity as a whole.

The logarithmic strategy ensured that groups containing few accessions, such as the Italian, Apple and Cherry strata, were represented in the sub-collection, and also that those groups with a large number of accessions contributed with relatively less accessions to the core collection. Logarithmic sampling therefore increases the probability of capturing the less frequent alleles compared to random sampling (BROWN, 1989). This strategy avoids the excessive sampling of accessions from large strata and increases the number of accessions sampled in the smaller strata, reducing bias due to group size (OLIVEIRA *et al*., 2010).

Once established, each sub-collection was evaluated as to its representativeness, i.e. its capacity to retain the variability of the initial collection. The representativeness of the core collection for the purposes of conservation, means maintaining the genetic variability. When comparing mean values, amplitudes, frequencies and variances for specific characteristics among the different members of the core and initial collections, it is expected that the intervals remain similar, while mean values move toward the median, and variances increase in the core collection (VAN HINTUM *et al*., 2000).

Table 2 - Core sub-collections of the tomato from BGH-UFV, established at a sampling density of 20% and 30%, from integrated and different type data, by the method of logarithmic stratified sampling

	Stratum[2]	20% density	30% density
QUANT[1]	SAL	850, 1989, 2211, 989 and 1211	850, 1989, 2211, 989, 1211, 1991, 489 and 1282
	SC	994, 4053, 1020 and 1490	994, 4053, 1020, 1490, 4006 and 970
	ITA	4054	1990 and 4054
	AP	2216 and 2234	2216 and 2234
	CHE	980 and 1254	980 and 1254
MULT	SAL	489, 1987, 2203, 989 and 2213	489, 1987, 2203, 989, 2213, 1989, 2202 and 181
	SC	975, 4053, 1708 and 991	975, 4053, 1708, 991, 4035 and 1497
	ITA	1990	1990 and 4054
	AP	2216 and 2234	2216 and 2234
	CHE	980 and 1254	980 and 1254
MOL	SAL	181, 2211, 1282, 1987 and 1499	181, 2211, 1282, 1987, 1499, 997, 1985 and 1993
	SC	1490, 1498, 1497 and 4055	1490, 1498, 1497, 4055, 4035 and 994
	ITA	4054	1990 and 4054
	AP	2216 and 2234	2216 and 2234
	CHE	773 and 980	773 and 980
COD	SAL	489, 1993, 2202, 181 and 1987	489, 1993, 2202, 181, 1987, 850, 2208 and 1282
	SC	975, 4035, 1490 and 4309	975, 4035, 1490, 4309, 991 and 1497
	ITA	1990	1990 and 4054
	AP	2216 and 2234	2216 and 2234
	CHE	980 and 1254	980 and 1254
SUM	SAL	850, 2202, 1282, 181 and 1993	850, 2202, 1282, 181, 1993, 2208, 489 and 2211
	SC	991, 4035, 1490 and 975	991, 4035, 1490, 975, 4053 and 166
	ITA	1990	1990 and 4054
	AP	2216 and 2234	2216 and 2234
	CHE	980 and 1254	980 and 1254

[1]QUANT: sub-collection formed from quantitative data; MULT: sub-collection formed from multi-category data; MOL: sub-collection formed from molecular data; COD: sub-collection formed from the integration of quantitative, multi-category and molecular data by the encoding of quantitative into multi-category data; SUM: sub-collection formed from the integration of quantitative, multi-category and molecular data through algebraic matrix addition. [2] Stratum: SAL - Saladinha; SC - Santa Cruz; ITA - Italian or Saladette; AP - Apple or Persimmon; CHE - Cherry

At the sampling density of 20%, the sub-collections COD-20 and SUM-20, obtained by data integration, stood out, as they displayed values for the amplitude coincidence index (AC) that were greater than 0.80 in relation to all the groups of characteristics, whether quantitative, multi-category or molecular (Table 3).

According to Hu, Zhu and Xu (2000) and Wang *et al*. (2007), a sub-collection is representative when its amplitude coincidence index is at least 80%. Sub-collections established from each set of characteristics, QUANT-20 MULT-20 or MOL-20, were therefore not considered representative, as they presented values for AC of 0.79, 0.77 and 0.78 for the molecular, quantitative and multi-category groups of characteristics respectively.

The variance of a sub-collection for each group of characteristics is another parameter that should be taken into account when assessing the representativeness of the sub-collections. By maintaining the amplitude of the characteristics, an increase in variance is expected in the sub-collections relative to the initial collection, since the number of individuals sampled is smaller. However, a significant decrease was seen in variance for the sub-collections, MULT-20 and MOL-20, in relation to the group of quantitative characteristics (Table 3), indicating that individuals with extreme phenotypic values were not included in the sampling process for these sub-collections.

For the sub-collection, QUANT-20, a significant increase was seen in variance for the group of quantitative

Table 3 - Amplitude coincidence index (AC) and mean variance of characteristics in the sub-collections formed at a sampling density of 20% of the data

Sub-collection[1]	Quantitative[2]	Multi-category[2]	Molecular[2]
QUANT-20	0.90 (204325.8**)[3]	0.80 (1.6508^{ns})	0.79 (0.1106*)
MULT-20	0.77 (119051.5*)	0.82 (2.0951**)	0.81 (0.1135**)
MOL-20	0.85 (120873.5*)	0.78 (1.5764^{ns})	0.92 (0.1431**)
COD-20	0.86 (138148.4^{ns})	0.85 (1.9171^{ns})	0.90 (0.1359**)
SUM-20	0.91 (144725.7^{ns})	0.83 (1.9431*)	0.88 (0.1307**)
IC	(144671)	(1.6938)	(0.0987)

[1]QUANT-20: sub-collection formed from quantitative data; MULT-20: sub-collection formed from multi-category data; MOL-20: sub-collection formed from molecular data; COD-20: sub-collection formed from the integration of quantitative, multi-category and molecular data by the encoding of quantitative into multi-category data; SUM-20: sub-collection formed from the integration of quantitative, multi-category and molecular data through algebraic matrix addition; IC: initial collection. [2] Corresponds to the group of characteristics used to estimate AC and variance. [3] Parenthesis: variance values; * ** significant at 5% and 1% probability, ^{ns} not significant at 5% probability by F-max test

characteristics. As this strategy gave a high value for AC, it can be inferred that this increase in variance is due to the greater frequency of extreme classes in this sub-collection, and furthermore, that the intermediate classes were under-represented. Consequently, the evaluation of AC and variance should be complementary when validating and choosing the best strategy for obtaining core sub-collections, allowing the conclusion that for sub-collections with the same value for CA, those with less variance are the most representative. Considering the values of AC and variance for the quantitative characteristics, the sub-collections, COD-20 and SUM-20, were the most efficient.

For multi-category characteristics, where the intermediate classes should also be represented in the sub-collections, the considerations relative to AC and the change in variance are the same as discussed above for quantitative characteristics. It is desirable to have sub-collections with a high amplitude coincidence index, together with a variance that is non-significant compared to the initial collection.

Regarding the group of molecular characteristics however, the core collections should be established in such a way as to preserve the greatest number of alleles and increase the frequency of rare alleles (alleles with a frequency of less than 5%). The ideal core collection is one that can maintain for a given locus the highest number of alleles of the same frequency. In the case of dominant molecular markers such as ISSR, the most suitable sub-collection is that in which the frequency of individuals is the same, possessing a mark or not. Thus, the greater its variance, the more efficient will be the sub-collection.

Considering the three groups of characteristics, quantitative, multi-category and molecular, the COD-20 strategy stood out in relation to SUM-20, with high values for AC, lower and non-significant estimates of variance for the quantitative and multi-category groups of characteristics, and greater variance relative to the molecular characteristics.

Where the sampling density was 30%, all the sub-collections presented values for AC of more than 80% (Table 4). However, QUANT-30 and MULT-30 gave large

Table 4 - Amplitude coincidence index (AC) and variance for the sub-collections formed at a sampling density of 30% of the data

Sub-collection[1]	Quantitative[2]	Multi-category[2]	Molecular[2]
QUANT-30	0.94 (213747.5**)[3]	0.85 (1.7624^{ns})	0.83 (0.1020^{ns})
MULT-30	0.89 (163845.9^{ns})	0.90 (1.9856*)	0.94 (0.1131**)
MOL-30	0.92 (158051.7^{ns})	0.85 (1.6108^{ns})	0.98 (0.1334**)
COD-30	0.92 (150664.5^{ns})	0.86 (1.8708^{ns})	0.94 (0.1279**)
SUM-30	0.91 (172489.8*)	0.85 (1.8032^{ns})	0.90 (0.1219**)
IC	(144671)	(1.6938)	(0.0987)

[1]QUANT-30: sub-collection formed from quantitative data; MULT-30: sub-collection formed from multi-category data; MOL-30: sub-collection formed from molecular data; COD-30: sub-collection formed from the integration of quantitative, multi-category and molecular data by the encoding of quantitative into multi-category data; SUM-30: sub-collection formed from the integration of quantitative, multi-category and molecular data through algebraic matrix addition; IC: initial collection. [2] Corresponds to the group of characteristics used to estimate AC and variance. [3] Parenthesis: variance values; * ** significant at 5% and 1% probability, ^{ns} not significant at 5% probability by F-max test

and significant estimates for variance in relation to the groups of quantitative and multi-category characteristics respectively, being therefore considered less efficient.

Only the sub-collections, MOL-30 and COD-30, were representative, with values for AC over 80%, together with variance values of the appropriate magnitude for the three groups of characteristics. This result demonstrates that at this sampling density, only molecular characterisation was enough to establish a core sub-collection as efficient as COD-30 in maintaining the maximum of germplasm variability.

For the graphical analysis of the retention of variability, encoding the quantitative and multi-category data into binary data, together with the molecular data, resulted in 288 classes, with 63 for the quantitative data, 173 for the multi-category data and 52 for the molecular data.

Figure 1 shows the frequencies of those classes relating to the quantitative, multi-category and molecular characteristics respectively, for the COD-20 sub-collection.

In each graph, for each class of characteristics, the frequency of accessions belonging to that class are shown, both for the initial collection and the sub-collection, with their deviation. It can be seen that only two of the quantitative classes were not retained in the sub-collection (indicated by arrows), i.e. none of the accessions chosen to make up COD-20 had that characteristic (Figure 1A).

For the multi-category characteristics, 142 of the 173 classes were represented in the sub-collection, that is, 82% of the phenotypes were retained, as they were present in at least one of the accessions that comprised COD-20 (Figure 1B). For the molecular data, variability was maintained for 92% of the loci, which represents the loss of five alleles (indicated by arrows) in the sampling process, either by fixing the presence of a mark or fixing the absence of a mark (Figure 1C). Accordingly, the retention of variability index (RVI) for this collection was 86.8%.

In the graphical analysis of the sub-collections, MOL-30 and COD-30, variability was retained for all classes of quantitative characteristics (Figure 2A), and no classes with a frequency of zero were seen. For the multi-category characteristics, the sub-collection, MOL-30, was less efficient than COD-30, as more classes of zero frequency were noted in MOL-30, where 29 of the 173 characteristics present in the initial collection were not represented, whereas in COD-30, the number of characteristics not sampled was 25 (Figure 2B).

Comparing the graphs of variation in allele frequency (classes of molecular characteristics) for MOL-30 and COD-30, a loss of one and three alleles was seen respectively for these sub-collections (Figure 2C). Dealing in this case with dominant molecular markers, the loss of alleles was recorded by fixing the presence or the absence of a mark (highlighted by arrows).

Figure 1 - Frequency variation for each class of characteristics, quantitative (A) multi-category (B) and molecular (C). Shaded circles indicate frequency for classes in the initial collection, empty circles indicate frequency for classes in the established core sub-collection, COD-20, (integration of quantitative, multi-category and molecular data by the encoding of quantitative into multi-category data, at a sampling density of 20%). Arrows indicate classes of characteristics with zero frequency in the sub-collection (A) and alleles that were fixed, for presenting either zero or maximum frequency in the sub-collection (C)

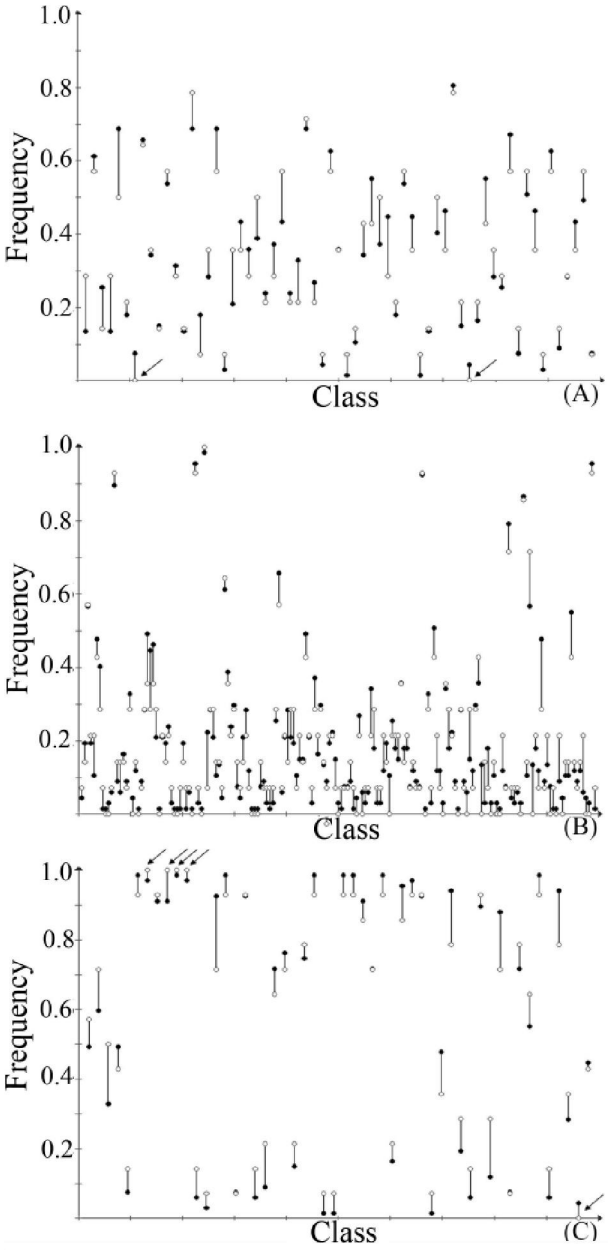
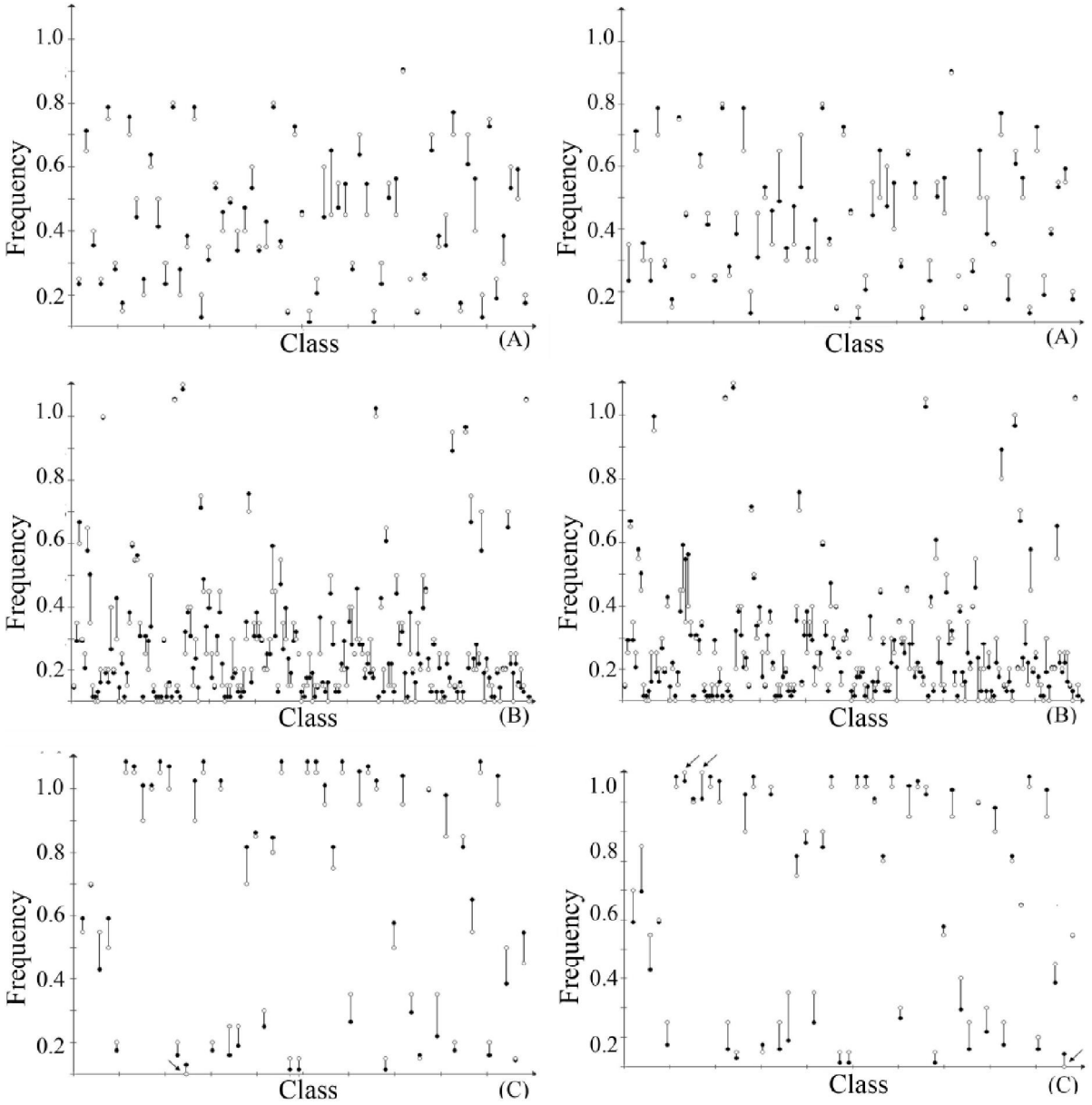


Figure 2 - Graphical analysis of the retention of variability for the sub-collections, COD-30, (on the left) and MOL-30 (on the right). Shaded circles indicate frequency for classes in the initial collection; empty circles indicate frequency for classes in the sub-collections. Frequency variation between the initial collection and the sub-collections, COD-30 and MOL-30, for each class of characteristics, quantitative (A) multi-category (B) and molecular (C). Arrows indicate classes of characteristics with zero frequency and alleles that were fixed, for presenting either zero or maximum frequency in the sub-collection



In general, from the graphical analysis it was concluded that, although the amplitude coincidence index and the mean of the variances of the characteristics in the sub-collections, MOL-30 and COD-30, demonstrate that both are equally efficient, COD-30 was slightly more so, as it presented an RVI of 93.75%, retaining a greater number of classes of characteristics in comparison with the sub-collection, MOL-30 (RVI = 89.6%).

A core collection is defined as set of accessions from a sample of germplasm, chosen to represent the maximum genetic variability of the initial collection with the minimum of redundancy (BROWN, 1989; FRANKEL, 1984). Whenever data of different types are available therefore, the establishment of core collections from the integration of such data should be prioritised. In this context, encoding quantitative into multi-category

data using the DEA-3 strategy proposed by Martins *et al*. (2011) was the most appropriate.

CONCLUSIONS

1. The evaluation of different sub-collections as to sampling density and data type, reveals very different collections, and that only the accessions 980 and 2234, belonging to the Cherry and Apple strata respectively, were present in all the sub-collections formed;
2. Only those sub-collections based on the integration of data were validated for all data sets at 20% density, the sub-collection, COD-20, being recommended for presenting more appropriate variance values and indices of amplitude coincidence;
3. At a density of 30%, the sub-collection, COD-30, proves to be the most representative, since as well as the appropriate amplitude coefficient and variance values, graphical analysis of the variability, followed by estimation of the retention of variability index, shows a value of 93.75%;
4. Whenever data of different types are available, the establishment of core collections from the integration of such data should be prioritised. In this context, the DEA-3 strategy proposed by Martins *et al.* (2011) was the most appropriate.

REFERENCES

ABADIE, T. *et al.* Construção de uma coleção nuclear de arroz para o Brasil. **Pesquisa Agropecuária Brasileira**, v. 40, n. 2, p. 129-136, 2005.

AGUILERA, J. G. *et al*. Genetic variability by ISSR markers in tomato (*Solanum lycopersicon* Mill.). **Revista Brasileira de Ciências Agrárias**, v.6, n. 2, p. 243-252, 2011.

BROWN, A. H. D. The case for core collections. *In*: BROWN, A. H. D.; FRANKEL, O. H.; MARSHALL, D. R.; WILLIANS, J. T. **The use of plant genetic resources**. Cambridge: Cambridge University Press: IPGRI, 1989, p. 136-156.

DWIVEDIL, S. L.; UPADHYAYA, H. D.; HEDGE, D. M. Development of core collection using geographic information and morphological descriptors in safflower (*Carthamus tinctorius* L.) germplasm. **Genetic Resources and Crop Evolution**, v. 52, n. 7, p. 821-830, 2005.

FRANKEL, O. H. Genetic perspectives of germplasm conservation. *In*: ARBER, W. K. *et al*. **Genetic manipulation**: **impact on man and society**. Cambridge: Cambridge University Press, 1984. p. 161-170.

GOMES, C. N. **Caracterização morfo-agronômica e diversidade genética em mandioca *Manihot esculenta* Crantz**. 2007. 82 f. Dissertação (Mestrado em Agronomia/Fitotecnia) - Universidade Federal de Lavras, Lavras, 2007.

HAO, C. Y. *et al*. Genetic diversity and core collection evaluations in common wheat germplasm from the Northwestern Spring Wheat Region in China. **Molecular Breeding**, v. 17, n. 1, p. 69-77, 2006.

HU, J.; ZHU, J.; XU, H. M. Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops. **Theoretical and Applied Genetics**, v. 101, n. 1-2, p. 264-268, 2000.

IPGRI, **Descriptors for tomato** (*Lycopersicon ssp*). Roma, Itália: International Plant Genetic Resources Institute, 1996. 56 p.

JANSEN, J.; VAN HINTUM, T. Genetic distance sampling: a novel sampling method for obtaining core collections using genetic distances with an application to cultivated lettuce. **Theoretical and Applied Genetics**, v. 114, n. 3, p. 421-428, 2007.

LI, Y. *et al*. Establishment of a core collection for maize germplasm preserved in Chinese National Genebank using geographic distribution and characterization data. **Genetic Resources and Crop Evolution**, v. 51, n. 8, p. 845-852, 2004.

MARTINS, F. A. *et al*. Integração de dados em estudos de diversidade genética de tomateiro. **Pesquisa Agropecuária Brasileira**, v.46, n. 11, p.1496-1502, 2011.

OLIVEIRA, M. F. *et al*. Establishing a soybean germplasm core collection. **Field Crops Research**, v. 119, n. 2-3, p. 277-289, 2010.

SNEDECOR, G. W.; COCHRAN, W. G. **Statistical Methods**. 7.ed. Ames: Iowa State University, 1980. 507 p.

UPADHYAYA, H. D. *et al*. Phenotypic diversity in the pigeonpea (*Cajanus cajan*) core collection. **Genetic Resources Crop Evolution**, v. 54, n. 6, p. 1167-1184, 2007.

VAN HINTUM, T. J. L. *et al*. **Core collections of plant genetic resources**. Roma: IPGRI Technical Bulletin, 2000.

VASCONCELOS, E. S. *et al*. Estratégias de amostragem e estabelecimento de coleções nucleares. **Pesquisa Agropecuária Brasileira**, v. 42, n. 4, p. 507-514, 2007.

VASCONCELOS, E. S. *et al*. Tamanho de coleção original, métodos de agrupamento e amostragem para obtenção de coleção nuclear de germoplasma. **Pesquisa Agropecuária Brasileira**, v. 45, n. 12, p. 1448-1455, 2010.

WANG, J. *et al*. A strategy on constructing core collections by least distance stepwise sampling. **Theoretical and Applied Genetics,** v. 115, n. 1, p. 1-8, 2007.

WANG, L. *et al*. Establishment of Chinese soybean (*Glycine max*) core collections with agronomic traits and SSR markers. **Euphytica**, v. 151, n. 2, p. 215-223, 2006.

WANG, Y. *et al*. Construction and evaluation of a primary core collection of apricot germplasm in China. **Scientia Horticulturae**, v. 128, n. 3, p. 311-319, 2011.

XU, H. M. *et al*. Sampling a core collection of island cotton (*Gossypium barbadense* L.) based on the genotypic values of fiber traits. **Genetic Resource Crop Evolution**, v. 53, n. 3, p. 515-521, 2006.

ZEWDIE, Y.; TONG, N.; BOSLAND, P. Establishing a core collection of *Capsicum* using a cluster analysis with enlightened selection of accessions. **Genetic Resources and Crop Evolution**, v. 51, n. 2, p. 147-151, 2004.