ORIGINAL ARTICLE

Maximization of minority classes in core collections designed for association studies

Alba María Vargas[1] · María Teresa de Andrés[1] · Javier Ibáñez[2]

Received: 8 June 2015 / Revised: 28 February 2016 / Accepted: 16 March 2016 / Published online: 23 March 2016

Abstract Core collections are nowadays widely employed in diverse studies on plant genetics. The more extensively used method to build core collections (maximization strategy) is based on the selection, from a global collection, of those accessions which maximize the number of different alleles and phenotypic classes (classes' richness). However, different core collections should be created for different types of studies, and though several years ago most of core collections were developed to make the characterization and use of germplasm collections easier with a smaller sample size, for either conservation or breeding purposes, today, they are widely employed for association studies that are broadly applied in plant genetic improvement. Following the M strategy, some alleles or phenotypic classes often appear in a very low frequency, which may reduce the power of the analysis, avoiding the detection of real associations (false negatives). In this work, we propose and evaluate a new way to build core collections using the maximization strategy in several sequential steps, to maximize the frequency of minority classes, thus increasing the statistical power of the association study.

Communicated by D. Grattapaglia

Electronic supplementary material The online version of this article (doi:10.1007/s11295-016-0988-9) contains supplementary material, which is available to authorized users.

✉ Javier Ibáñez
javier.ibanez@icvv.es

[1] Instituto Madrileño de Investigación y Desarrollo Rural, Agrario y Alimentario (IMIDRA). Finca "El Encín", Ctra. A2, Km 38.200, Alcalá de Henares, Madrid 28800, Spain

[2] Instituto de Ciencias de la Vid y del Vino—ICVV, CSIC, Universidad de La Rioja, Gobierno de La Rioja, Carretera de Burgos Km. 6, Finca La Grajera, Logroño 26071, Spain

Keywords Association mapping · Germplasm collection · M strategy · Minor classes · Statistic power

Introduction

The maintenance of plant species in germplasm collections is of great importance to conserve their genetic diversity, which in many cases has suffered deep genetic erosion due to different social and agriculture factors, such as the importation of more productive foreign varieties. Given the high number of accessions integrating some collections, it is difficult to characterize them completely, so the establishment of smaller collections representing the genetic diversity of the large collection is very useful to carry out certain studies. Such collections are called core collections (CCs) and contain a much lower number of accessions than the initial collection; so most of the published CCs contain between 5 and 20 % of the global collections (van Hintum et al. 2003).

Since Frankel (1984) proposed the concept of core collection, a body of literature on the theory and practice of core collections has accumulated. Many approaches for selection and evaluation of CCs have been proposed and used (see Odong et al. (2013)). The different methods to build CCs depend on several factors, and the level of stratification of the collection is among the most important ones. In stratified collections criteria like the size of each group (Brown 1989), the phenotypic Gower distance (D strategy) (Franco et al. 2005) or the genetic diversity in each group (H strategy) (Schoen and Brown 1993) have been used. The selection of accessions for the core collection in non-stratified collections (or in each group of stratified collections) may be at random (R strategy), but nowadays, the allelic maximization is the more used criterion (M strategy) (Schoen and Brown 1993). The M strategy examines all possible core collections and singles out those that maximize the number of observed different classes (richness) at the markers used, which may be molecular, morphological, phenological, etc. (Gouesnard et al. 2001). Bataillon et al. (1996) observed that M strategy is able to capture the allelic richness both in neutral and selected loci, which usually present large frequency differences. This method not only defines the number of accessions needed to capture the desired variability but also identifies each individual accession that must be included in the CC. From a practical point of view, searching throughout all possible core collections is unfeasible when the collection to be sampled is large, because the number of possible combinations grows factorially with the sizes of the core and of the whole collection (Gouesnard et al. 2001). For that reason, the software used for establishing core collections implement algorithms to improve results while keeping the calculation time needed in reasonable terms. Because of this, and especially if the whole collection is very redundant, different runs of a software may provide different outcomes regarding the composition of the core collection and the frequency distribution of the classes within each trait.

Besides, it is important to take into account the purpose of the core collection, since different phenotypic distributions could be required for different applications, and so, distinct methods for building them should be used. So far, many CCs have been built in different ways for conservation or breeding purposes, but currently, genetic association studies are proliferating (Khan and Korban 2012), and they have specific requirements.

Association mapping, or linkage disequilibrium (LD) mapping, has been applied in numerous plant species and essentially involves searching for genotype-phenotype correlations in a population, which is commonly a collection of individuals. Compared with traditional quantitative trait loci (QTL) linkage analysis, association mapping has several advantages: It provides better genome coverage of marker polymorphisms than any biparental population, it has a higher mapping resolution and the possibility to detect multiple allelic effects, and it does not require the development of segregating populations. Several research groups have successfully performed association analyses of multiple traits using core collections (Bordes et al. 2013; Carpio et al. 2011; Fernandez et al. 2014; Holbrook and Anderson 1995; Kwon et al. 2012; Li et al. 2011; Soto-Cerda et al. 2013, 2014; Upadhyaya et al. 2012, 2013; Vargas et al. 2013a, b; Wang et al. 2011; Zhao et al. 2010; Zorić et al. 2012).

Association studies in plants show generally problems related to the collection used. The main problem is genetic relatedness among individuals in the collection. This may cause a genotype-phenotype covariance, and many genetic markers across the genome will appear to be associated with the phenotype, when in fact, these genetic markers simply capture the genetic relatedness among individuals (Myles et al. 2009). The use of mixed linear models (MLMs) that correct for relatedness through incorporating population structure and/or kinship information has reduced the rate of false positives (Weir 2010; Zhang et al. 2010).

Another problem of association studies are the false negatives, related to the power of the association test. The statistical power for identifying markers associated with quantitative traits depends on several factors, including heritability, the number of causal variants, and their frequency (Shin and Lee 2015). For instance, Gonzalez-Martinez et al. (2008) observed that the power to detect association decayed rapidly for low frequency ($0.1 < \mathrm{MAF} < 0.2$) and rare alleles ($\mathrm{MAF} < 0.1$). Whitt and Buckler (2003) recommended excluding polymorphisms with a frequency less than 5 % from the analyses, partly because there is rarely enough statistical power to test for association at these low frequency polymorphisms. This problem may be even raised in complex traits associated with population structure, because the correction of structure and kinship in MLM also weakens the real associations (Atwell et al. 2010; Liu et al. 2016). Although this issue is normally considered regarding the genotypic part, it affects in the same way the phenotypic one. Certain classes of interesting agronomical or commercial traits are represented in low frequency in the global collections, and traditional core collections built over the basis of maximizing the total diversity can decrease even more the frequency of these minority classes. In this work, we propose a new approach based on the M strategy to obtain core collections more suitable for association mapping, which could improve the power of these analyses to detect associations by reducing false negative results for traits with classes at low frequency.

Material and methods

Three different methods for building core collections were tested using two datasets of two plant species (rice and grapevine) for which traits of different characteristics had been described.

Rice data

A phenotypic dataset for 1000 rice accessions available at http://www.genebank.go.kr/eng/PowerCore/ (Kim et al. 2007) was used. It contains values for 39 phenotypic traits (28 qualitative and 11 quantitative traits) (Online Resource 1).

Grapevine data

The grapevine plant material consisted in 246 unique genotypes of a table grape collection maintained at the germplasm bank of El Encín (IMIDRA, Spain), which was characterized for 37 morphological and phenological descriptors (17 qualitative and 20 quantitative traits) and for 20 nuclear simple sequence repeats (SSRs) (Ibáñez et al. 2009) (Online Resource 2).

Generation of core collections

Core collections were generated with the software MStrat 4.1 (Gouesnard et al. 2001). This software considers each value of a qualitative variable or marker allele as an individual class, while values of quantitative variables are binned into a series of discrete classes of identical size along the variable range of variation, and which number is chosen by the user. We split each quantitative variable into ten classes (except for a few of them, with a limited range of variation, for which we used five classes). This arbitrary number was chosen to give the quantitative variables a similar weight to that of a microsatellite marker (10.1 alleles in average in the grapevine set) and to minimize the number of classes appearing only once in the global collection.

Firstly, redundancy of each collection was examined to know the number of accessions necessary to represent a high percentage of the total collection. Random (R) and maximization (M) methods were used. In order to allow for comparisons between the different methods in the two species, the sizes of the final CCs were not determined from the redundant data, but arbitrarily fixed at 96 entries (the term “accessions” refers to elements that constitute the whole collection and “entries” are elements of the core collection).

Three core collections of 96 entries were built using the Maximization strategy (Schoen and Brown 1993), implemented in MStrat 4.1 (Gouesnard et al. 2001). The software was set to use 100 iterations to generate each of 100 potential core collections (CC replicates) in every run. The primary criterion used to select any CC was classes’ richness, defined as the sum, across all the variables or markers, of the number of different classes or alleles represented among the entries. In case of tie between two CC replicates for richness, Shannon Index was used as secondary criterion of optimization or selection. Each CC was built according to the three following methods, two of which comprise different sequential steps:

Method 1 Core collection obtained in one step. Procedure: *Generation of 100 CC replicates of 96 entries with MStrat → Selection of the CC with the highest classes’ richness, and, in case of ties, with the highest Shannon Index (CC96).*

Method 2 Three subcollections, each one obtained in one step. Procedure: *Generation of 100 CC replicates of 32 entries with MStrat → Selection of the subcollection with the highest classes’ richness, and, in case of ties, with the highest Shannon Index → Removal of the accessions included in this subcollection from the total collection and repetition of the process twice more → Union of the three core subcollections of 32 entries to obtain a CC of 96 entries (CC32Sx3).*

Method 3 Three subcollections, each one obtained in four steps. Procedure: *Generation of 100 CC replicates of 32 entries → Selection of the accessions more frequently represented in the 100 replicates and set them in the kernel (compulsory entries) for the next step → Repetition of the process twice more (the objective is to set the accessions more represented in each 100 replicates along three steps, trying to avoid accessions that could be included in the core collection by random) → Generation of 100 new CC replicates of 32 entries, all of which will include the accessions set in the kernel in the three previous steps → Selection of the subcollection with the highest classes’ richness, and, in case of ties, with the highest Shannon Index → Removal of the accessions included in this subcollection from the total collection and repetition of the whole process twice more → Union of the three core subcollections of 32 entries to obtain a CC of 96 entries (CC32Fx3).*

Evaluation of core collections

The three core collections built from each dataset with different methods were compared and their quality was evaluated according to the searched goal. CCs generated for both datasets were analyzed for Shannon Index and total and minority classes’ frequencies. Besides, a genetic structure analysis was done on grapevine data, for which microsatellite data were available, to evaluate the possible differences in genetic structures obtained by the different methods proposed. It was done using 19 non-linked SSRs (all those included in Online Resource 2, but VVMD25), and software Structure 2.3.2.1 (Pritchard et al. 2000). A model with a putative number between one and ten populations and correlated allele frequencies (Falush et al. 2003) was assumed. Monte Carlo Markov Chain run-length period of 1,500,000, with 500,000 burn-in steps, and 10 iterations for each number of putative populations were used. Evanno criterion (ΔK) was used to decide the number of populations (Evanno et al. 2005). A threshold of 0.8 for the membership coefficient (Q) was fixed for population assignment.

In addition, in order to evaluate the effect of each method in association analysis, specifically in the rate of false negatives (Type II error), simulations were done in the grapevine global collection to test against the phenology trait “Time of Budburst.” Data was simulated to create an association between the class “very late” (numerically 9, Online Resource 2) and the recessive allele (C) of a biallelic locus with complete dominance (TT and CT with no effect on the trait). Genotypic data was generated with a priori frequency of 0.325 for allele C. Genotypes were first assigned to the accessions presenting class 9, considering three degrees of linkage disequilibrium: correlations between class 9 and genotype CC of 1.0, 0.8, and 0.6 respectively (Online Resource 3). Then, genotypes were assigned to the remaining accessions using the expected frequencies at Hardy-Weinberg equilibrium as assignment probabilities. Data for 5 different markers were simulated for a priori correlations of 0.6 and 0.8 and 1 marker for total correlation, and all these 11 markers were used in association analyses in the global and the 3 core collections.

Association analyses were done using a mixed linear model (MLM) implemented in TASSEL v3.0, applying an optimum level of compression and the P3D approach for variance component estimation (Zhang et al. 2010). A Q matrix for each collection was obtained with Structure 2.3.2.1 as detailed above and kinship coefficients were calculated using the same 19 non-linked SSRs used in Structure and the estimator of Ritland (1996) implemented in SPAGeDi 1.3 (Hardy and Vekemans 2002).

Results

Three methods were used for building core collections on two different datasets with the objective of establishing a new approach to maximize in the core collections the representation of minority classes, what we hypothesize that would improve association analyses through reducing false negatives. MStrat software, developed by Gouesnard et al. (2001), was used to evaluate the redundancy of datasets and to build the core collections.

Redundancy in the collections and efficiency of M strategy

The rice dataset contained 1000 accessions and 182 different classes, while grapevine dataset included 246 accessions and 461 classes (Online Resources 1 and 2). The maximum numbers of classes were 191 (rice) and 464 classes (grapevine), as determined by the number of classes in the qualitative variables and molecular marker alleles, and the number of classes arbitrarily established for the quantitative variables. The redundancy analysis determined that 37 accessions (4 % of the total number) and 54 accessions (22 %) were enough to account for the total classes' richness existing in rice and grapevine datasets, respectively. A core collection of 32 entries represents up to 97.2 % of the diversity existing in the rice whole collection and 93.9 % in the case of grapevine.

The variability captured by the M strategy was, on average, 32 % superior to that obtained using the R strategy (random selection of accessions) in rice, while in grapevine, collections obtained using M strategy contained 14 % more classes than those obtained using R strategy.

Generation of the core collections

The three core collections of 96 entries (CC96, CC32Sx3, and CC32Fx3) obtained with the three different methods (1, 2, and 3 respectively) are shown in Online Resource 1 (rice) and Online Resource 2 (grapevine).

All the 100 CC replicates obtained with Method 1 included all the classes existing in the whole collections (maximum richness), and so the selection of CC96 was based on the Shannon Index. In both species, rice and grapevine, only one collection replicate was obtained with the highest Shannon index and was selected. This was also the case of the 3 subcollections of 32 entries (CC32S) obtained using Method 2 with the 2 species data.

Each CC32F was obtained in four steps with Method 3. The first three steps were used to select those accessions which appeared more frequently in the 100 replicates obtained with MStrat software. The selected accessions were set in the kernel file, and so, they were always included in the subsequent collection replicates obtained. In the second and third steps, new accessions were selected on the same basis and added to the kernel file. The number of selected entries included in the kernel in each step varied between 4 and 11, depending on the frequency of the more repeated accessions (Online Resource 4). So, in the case of rice data, for the building of the first core collection of 32 entries, the first kernel included accessions appearing in 100 % of replicates, the second kernel was formed by accessions appearing in at least 95 % of replicates, and 82 % was the threshold frequency for the inclusion in the third kernel. For the successive core subcollections of 32 entries, those percentages were lower (Online Resource 4). For grapevine data, the percentages used to include entries in the kernel were lower than for rice in every step: for instance, 82, 71, and 61 % in the first core subcollection of 32 entries (Online Resource 4). The fourth step consisted in the selection of the corresponding CC32F based on the highest richness, and, if needed, on the highest Shannon Index. Using these two criteria, only one of the collection replicates was obtained for each of the three CC32F in rice. In grapevine, two replicates showed identical number of classes and Shannon Index in the generation of the first subcollection and differed in only two entries. One replicate was chosen, and it was noted that the two excluded accessions were incorporated in the following subcollection. For the second and third grapevine subcollections, only one replicate was obtained with the maximum richness and Shannon index.

Composition of the core collections

As regards the entries shared by the different core collections obtained from rice data, only 38 entries appeared in the three CC (40 % of the 96 entries included) (Fig. 1). Forty entries (42 %) were common between CC96 and CC32Sx3 and

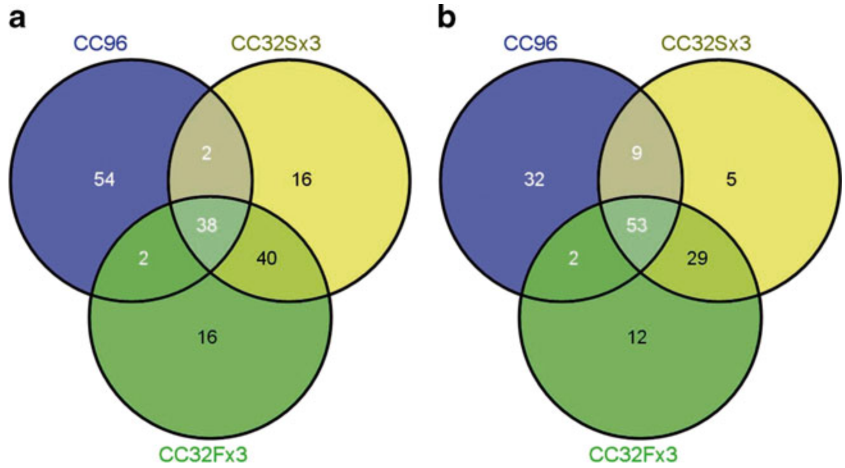

Fig. 1 Venn diagram showing the common entries in the three core collections: CC96 (Method 1), CC32Sx3 (Method 2), and CC32Fx3 (Method 3). **a** Rice collections; **b** grapevine collections

between CC96 and CC32Fx3, while 78 entries (81 %) were common in CC32Sx3 and CC32Fx3. In grapevine, 53 entries were present in the 3 core collections (55 %). The common entries between CC32Sx3 and CC32Fx3 were 82 (85 %), while they shared 65 and 57 % of their entries with CC96 (62 and 55 entries, respectively).

Evaluation of the diversity included in the core collections

The core collections obtained with Method 1 for both datasets accounted for 100 % of the classes existing in the whole collections, while the percentage of classes represented in the core collections built with methods 2 and 3 was always above 99 % (Table 1). In rice global collection, 23 unique classes were detected and distributed in 23 accessions; while in grapevine set, 32 unique classes were distributed in 24 accessions. The number of unique classes was raised in CC96 with respect to the whole collection by 82 (rice) or 112.5 % (grapevine). In the case of CC32Sx3 and CC32Fx3, the raise in the number of unique classes was much lower in rice (1 and 2 unique classes more respectively) and intermediate in grapevine (15 and 8 unique classes more, respectively, Table 1).

The frequency distribution of the classes was also different in the CC96 and the two CC32x3 (Fig. 2). The low frequency classes (left sides of the figures and small windows) showed a higher frequency in CC32x3. For instance, in rice, there were 128 classes present in at least 4 entries (absolute frequency ≥4) in both CC32x3 against 108 classes in CC96. On the contrary, CC96 were enriched in classes with high frequency (right sides of the figures). In grapevine, more evident differences were detected in the number of classes with absolute frequency ≥3, with 380 and 382 classes in CC32Sx3 and CC32Fx3, respectively, against 352 classes in CC96.

This different frequency distribution of classes in each core collection type was also revealed by Shannon Index (H). Large differences were found between grapevine and rice data used, but within each dataset, both CC32x3 showed similar values and higher than their corresponding CC96 (Table 1).

Minority or low frequency classes are especially important for the purpose of this work. Figure 3 shows the pairwise differences between the CCs in the number of classes with minimum absolute frequencies between 2 and 10. There were few differences in the number of classes when comparing CC32Sx3 and CC32Fx3, and these differences were positive or negative. Nevertheless, for these minority classes, there

Table 1 Description of diversity in the global and core collections obtained in rice and grapevine

		Global	CC96	CC32Sx3	CC32Fx3
Rice	Shannon Index	23.40	29.33	34.00	33.85
	Richness (total number of classes)	182	182	180	182
	Number of minority classes (freq < 10)	58	100	95	99
	Number of unique classes (freq = 1)	23	42	24	25
	Number of minority classes with freq < 4	39	74	52	54
Grapevine	Shannon Index	85.65	88.54	90.57	90.25
	Richness (total number of classes)	461	461	459	457
	Number of minority classes (freq < 10)	156	247	244	231
	Number of unique classes (freq = 1)	32	68	47	40
	Number of minority classes with freq < 4	81	136	111	117

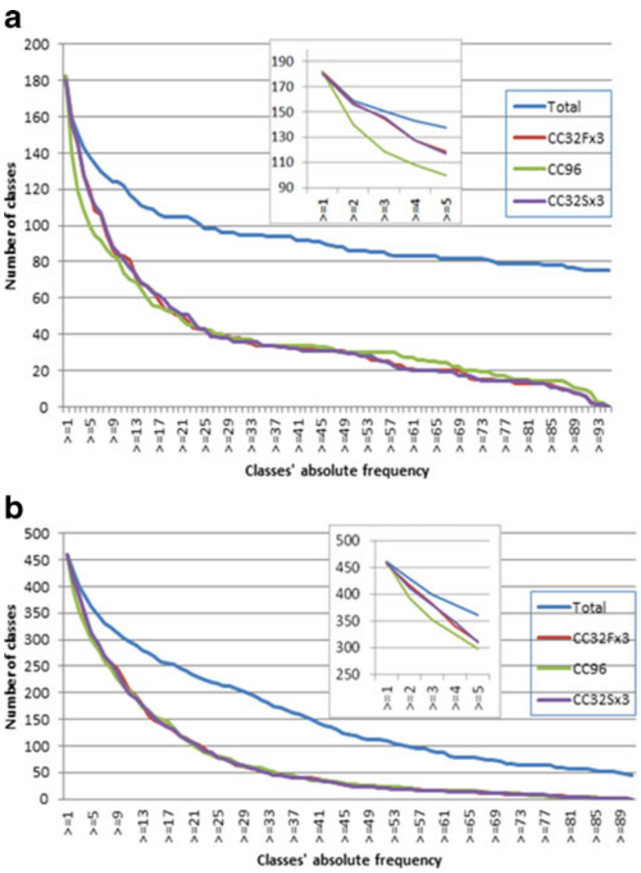

Fig. 2 Representation of the number of classes against absolute frequencies in the different core collections. The small windows are zoomed presentations of the classes with low absolute frequency. **a** Rice; **b** grapevine

were clear differences between any of the CC32x3 and CC96. In all the range considered, the difference was positive in favor of the CC32x3. For instance, in rice there were 20 classes more in both CC32x3 than in CC96 which were present in at least 4 entries.

Structure analysis in grapevine

Genetic structure was evaluated only in grapevine core collections since this dataset included molecular data. Results obtained for the CCs were compared with those observed in the global collection of 246 accessions (C246), for which two possible structures were obtained: one with two populations (C246-K2-Q1 and C246-K2-Q2) and another one with four populations (C246-K4-Q1 to Q4) (Online Resource 5a). These structures are related to the geographical origin and breeding history (*Vitis* International variety catalogue, www.vivc.de) of the varieties. The population Q1 is maintained in the two structures, while Q2 in C246-K2 was split into C246-K4-Q2 and C246-K4-Q3, and C246-K4-Q4 was formed essentially by accessions considered admixed (membership coefficient Q below 0.8) in the C246-K2 (Online Resources 5b, 6 and 7).

Regarding the CCs, a structure with two possible populations was obtained for CC96, while three possible populations were identified for both CC32x3 (Online Resource 5a). The CC96 structure was similar to C246-K2, with 59 entries in equivalent populations (Online Resources 5b, 6 and 7). The three CC32Sx3 populations essentially corresponded to Q1, Q2, and Q3 populations in C246-K4, with 38 entries in equivalent populations of C246-K4. In the case of CC32Fx3, the situation is similar, with 42 entries in equivalent populations of C246-K4. The frequencies of admixed entries in the structures obtained for CC32x3 were similar to that found in C246 and higher than the frequency obtained in CC96 (Online Resource 7), ranging from 17 % in CC96 to 45 % in CC32Sx3.

Association analysis on grapevine data

Biallelic data were simulated for 11 markers to test associations with the phenology trait Time of Budburst, where the class very late (numerically 9) had a frequency of 0.105 in the global collection, 0.063 in CC96, 0.115 in CC32Sx3, and 0.125 in CC32Fx3. Frequency for allele C was set a priori in

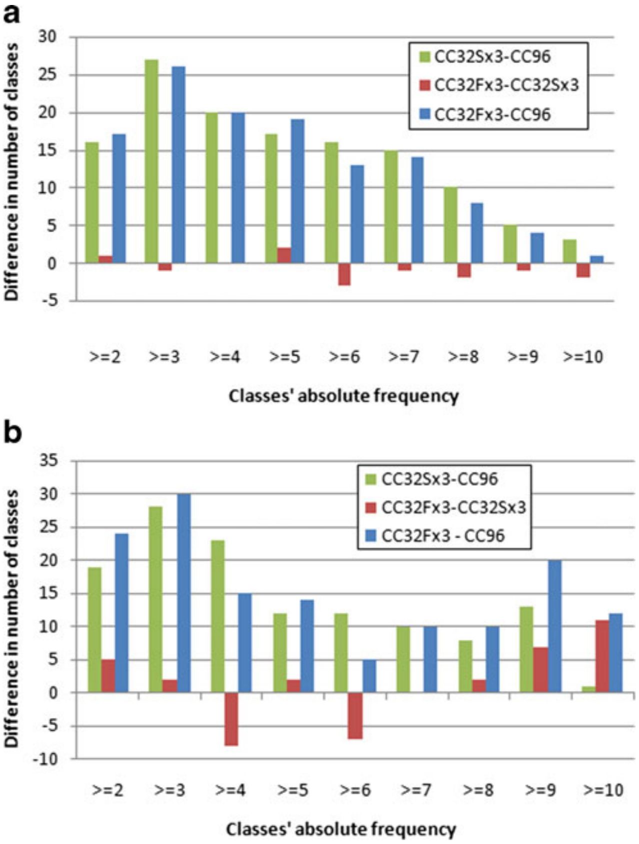

Fig. 3 Pair-wise differences between the three methods regarding the gain of minority classes. **a** Rice; **b** grapevine

0.325, corresponding to an expected CC genotypic frequency of 0.11 (in HWE), to allow testing for a total correlation with class 9. Data was simulated for different degrees of linkage disequilibrium, using a priori correlation coefficients between class 9 and genotype CC of 1.0 (SNP0), 0.8 (SNP1 to SNP5), and 0.6 (SNP6 to SNP10). Results of the simulation are shown in Online Resource 8 with the allele frequencies and correlations obtained for each marker in the global collection.

Association analyses were done for each of the markers with the MLM + Q + K model, using as cofactors the genetic structures described above (C246-K4 was used for the global collection). *P* values obtained for the phenology trait Time of Budburst were higher (less association) in CC96 than in the two CC32x3 in 8 of the 11 simulated markers (Fig. 4a), and only for SNP3, *P* value in CC96 was lower than in CC32Fx3. Using a significance threshold of 0.05, five different SNPs were not associated in CC96, and one in CC32Sx3, corresponding to false negatives, while all the SNPs were associated in CC32Fx3 (Fig. 4a). In all the five non-associated SNPs in CC96, the absolute frequency of entries with genotype CC and class 9 was below 5. Using a significance threshold of 0.01, seven SNPs were not associated in CC96, three in CC32Sx3, and two in CC32Fx3. The fraction of variance explained by the tested SNPs was also different in the different core collections. It was lower in CC96 than in the two CC32x3 for 7 of the 11 SNPs. Only for SNP3 it was higher in CC96 than in CC32Fx3 (Fig. 4b). The highest fraction of variance explained by the marker in CC96 was 0.19, lower than those found in CC32Sx3 (0.30) and in CC32Fx3 (0.31).

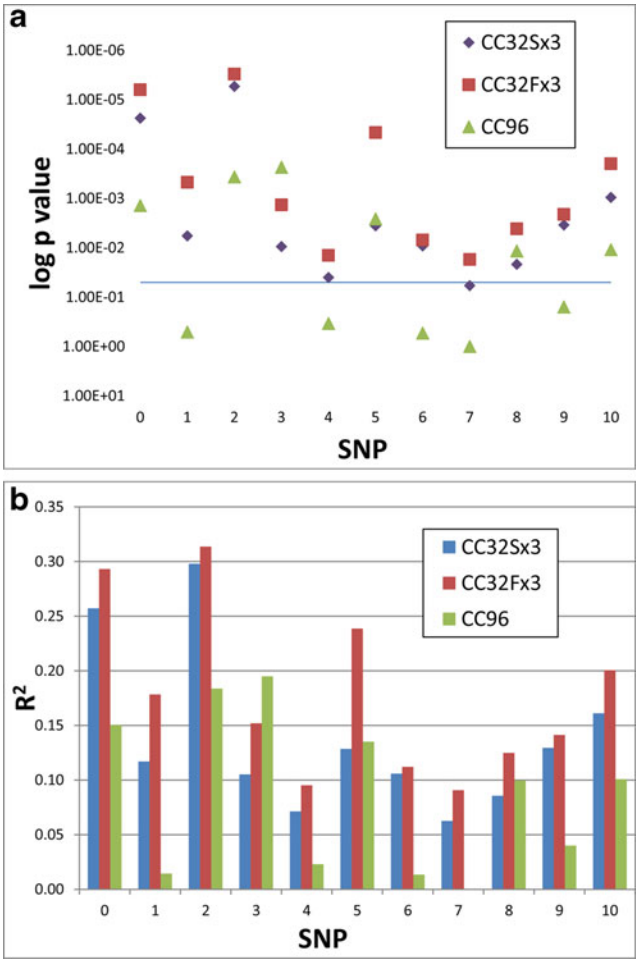

Fig. 4 **a** Plot of *P* values obtained in the association analyses of the trait "Time of Budburst" and each of the 11 SNPs with simulated data; *horizontal line* indicates the threshold for significance at 0.05. **b** Fraction of variance of the trait "Time of Budburst" explained by each of the 11 SNP markers with simulated data in the association analyses

Discussion

In wide germplasm collections, the cost of certain molecular approaches can be high and carrying out a complete phenotypic and molecular characterization may be very difficult, so it is necessary to reduce the number of the accessions to analyze while keeping the maximum amount of variability. This is normally done through the construction of core collections, a small representative sample of the whole collection. There are different methods that can be used to achieve this, depending on the purpose of the core collection and the availability of data collected in the whole collection. For association studies, the collection should be representative of the existing diversity for the traits under study, but moreover it is important that every phenotypic class is represented several times to increase the power of the test. This increased power contributes to diminish spurious associations and lack of association detection with low frequency data. Thus, for this type of studies, some redundancy is convenient, and so, a compromise is needed when constructing these core collections between minimizing the number of accessions and accounting for the maximum representation of the less frequent classes.

In this paper, a new approach was used to build core collections with the objective of increasing efficiency and maximizing the representativeness of minority classes/alleles. For this purpose, two alternative methods were used to build independently three subcollections that were then joined. The core collections obtained in this way (CC32Sx3 and CC32Fx3) were compared with another core collection (CC96) obtained using a standard one-step method (Le Cunff et al. 2008).

Collections redundancy

Redundancy evaluation constitutes a previous step to determine the number of entries needed to capture the required diversity, or, if the number of entries is predetermined (like in this case), to know the diversity that is possible to capture. The idea under the alternative approach proposed here, by creating the CC in three steps, is to have each class represented by a minimum of three entries, for most of the classes. This requires that the size of each subcollection is large enough to account for most of the richness of the whole collection. Thus, the redundancy analysis also helps to determine the number of entries to be included in each core subcollection. In the examples studied here, 32 entries included 97 % of total richness in rice and 94 % in grapevine. Ideally, each subcollection should include the highest possible richness, but below 100 %, to avoid unnecessary redundancy (new entries which do not contribute with additional classes). In this work, a number of 96 entries was predetermined for the core collections, because it is a number already used for association studies (Vargas et al. 2013a), it allows for a three-step method with a number of entries for each subcollection (32) that accounts for more than 90 % of classes' richness, and, from a practical point of view, it is appropriated to work with 96-well plates.

A large difference in redundancy was found between rice and grapevine because two very different datasets were chosen to illustrate the approach: rice dataset has a much larger number of accessions (1000 vs. 246) and a much smaller number of classes (182 vs. 461). This is probably influenced by the fact that rice is inbred, while grapevine is outbred, causing a very different partitioning of diversity. The 96 entries included in the core collections represented 9.6 % of the total number of accessions in rice and 39 % in grapevine.

Two methods were used to select accessions throughout the redundancy analysis: a random selection of accessions (R method) and a selection of accessions using the Maximization strategy. In both datasets, M (maximization) method was superior in gain of diversity compared to R (random) method. The results obtained here for grapevine data are similar to those obtained by McKhann et al. (2004), who observed a diversity gain of 10 % in *A. thaliana*, or by Le Cunff et al. (2008), who observed a diversity gain of 15 % in *V. vinifera*. Nevertheless, a gain of diversity more than twice higher (32 %) was obtained in rice with M against R method, probably because of the high redundancy existing in this collection. Unlike these results, Ronfort et al. (2006) did not detect a clear difference between both strategies in *Medicago truncatula*, maybe due to the low redundancy and linkage disequilibrium present in their collection.

Method comparison

The efficiency of the different strategies to establish core collections is evaluated in deep in several studies (Bataillon et al. 1996; Escribano et al. 2008; Franco et al. 2006; Gouesnard et al. 2001; Schoen and Brown 1993), and M strategy is generally the most efficient in the exclusion of redundancy. M strategy is nowadays widely used, and it is implemented in two publicly available programs, at least. In this work, MStrat (Gouesnard et al. 2001) was chosen because it allows to predetermine the number of entries to be included in the CC, while PowerCore (Kim et al. 2007) does not.

The different purposes of core collections make necessary additional approaches to obtain the more appropriated collection for each use. In this sense, different criteria to evaluate the quality of distinct types of core collections have been recently discussed (Odong et al. 2013). These authors stated that the choice of criteria for evaluating core collections depends on the objectives of the core and propose different ways according to the aim. In this work, we focused in a concrete application of core collections widely used nowadays: association studies, which was not considered in that work, and for that reason, we evaluated the obtained CCs in a distinct way. Our measure of quality is the increase in the frequency of minority classes, and the decrease in the number of unique classes in the resulting CCs, which we hypothesize that will improve the suitability of the CCs for this particular aim.

In this kind of studies, avoidance or reduction of spurious positive associations due to unique or minority classes could be an important save of time. Even more important is to avoid false negatives due to the low frequency of certain classes within interesting commercial traits, which could prevent detecting of real associations explaining certain variation of the trait.

Two methods to build core collections were used to try to increase the frequency of minority classes and were compared with another standard method (Method 1). Methods 2 and 3, based on the merge of three independent subcollections to constitute the definitive core collection, showed a clear decrease in the number of unique classes (useless for association analyses) and a raise in the frequency of minority classes compared to Method 1 (Table 1). In both datasets, the largest raise was found in the number of classes present in, at least, 3 entries (absolute frequencies ≥3); in rice, there were 27 more classes with this frequency in CC32Sx3 than in CC96 and in grapevine, there were 30 more classes in CC32Fx3 than in CC96.

Shannon Index accounts for both abundance and evenness of the classes present, and it was also higher in the collections CC32x3. Methods 2 and 3 against Method 1 worked better on rice data than on grapevine data, due to the higher redundancy (more accessions, less richness) existing in the rice dataset.

Method 3 consisted of fixing the accessions more frequently represented in 100 replicates along 3 steps, trying to favor those accessions that are more often included in the core subcollection. Their frequent presence indicates not only that they contribute with a high variability but also that they complement better with the diversity provided by other entries to maximize the richness of the core subcollection. In the case of the rice dataset, the frequency thresholds used to select entries for the kernel were higher than in grapevine in every step (Online Resource 4), because of the presence of certain entries that clearly contribute more than others to the core collection variability. The high heterozygosity and diversity of grapevine dataset allows for a larger number of possible combinations of accessions with high variability and Shannon Index, giving place to a larger diversity of entries in the core collection replicates. Thus, the threshold used in each case for the selection in the kernel was empirically established from the results obtained, aiming to include around eight entries in every step.

Unlike Method 3, Method 2 requires a single step for the construction of each subcollection. No clear differences were observed between methods 2 and 3, indicating that the two methods similarly favor both the representation of classes and the evenness of their frequencies. In fact, the number of common entries in both collections was high: 78 in rice and 82 in grapevine.

Structure evaluation in grapevine CCs

It is well known that other factors can influence association studies, such as population structure, parentage relationships, and linkage disequilibrium (Whitt and Buckler 2003), which are characteristic of each species, even each collection, and association mixed models may take them into account to reduce spurious results. In this work, the genetic structures obtained in the grapevine global collection and in the different corresponding CCs have been evaluated using a Bayesian algorithm.

The structures obtained for C246-K2 and CC96 are similar, as are those obtained for C246-K4 and CC32x3.These results are consistent with that previously published by Vargas et al. (2013a) and Vargas et al. (2013b), who obtained 2 populations for a collection of 96 table grape entries and 3 populations using a collection of 127 table grape entries, very similar to those obtained for CC96 and CC32x3, respectively. Bacilieri et al. (2013) used the same set of microsatellite markers for a genetic structure analysis in a large collection of 2096 *V. vinifera* genotypes and observed two possible structures of 3 and 5 populations. Among the five populations, three with table cultivars were included: (a) wine and table—Iberian Peninsula and Maghreb, (b) Table—East, and (c) Italy and Central Europe, that would be equivalent to the Q2, Q3, and Q1, respectively, detected in the two CC32x3. Though Bacilieri et al. (2013) suggested that K4 and K6 are not appropriate clustering levels for grapevine, it must be noted that only table grape accessions were included in this work. The similarity between the different structures obtained indicates that *V. vinifera* subsp. *sativa* has a well-defined stratification, which effects can be corrected in association studies.

CC96 showed the lowest proportion of admixed entries (Online Resource 7). Moreover, a significant raise was observed in the proportion of entries assigned to CC96-Q2, compared to C246-K2-Q2, indicating a strong bias in the representation of ancient cultivars. CC32x3 collections showed higher percentages of admixed entries, similar to C246, and a more evenly distributed number of entries in each population. These characteristics point out to a better distribution of entries in CC32x3 than in CC96 for association studies.

Here, the genetic structures have been evaluated after obtaining the CC, but the use of information about the existing stratification may also be considered before starting the building of the core collection. In case the available number of accessions in the global collection is large enough, the core collections may be created independently in each subpopulation using any of the alternative methods proposed here.

Association analysis

The examples tested on virtual data showed that the frequency of minority classes may be crucial for association analysis. Using biallelic data simulated for the grapevine collection, with different degrees of linkage disequilibrium between the marker and the trait, several clear examples of absence of association detection (false negatives) are provided. The results showed that the effect of the minority classes' frequency is more critical as the linkage disequilibrium is lower, and it affects both the *P* value (association detection) and the fraction of variance explained by the marker.

In base to the results obtained, a proposal is made for the construction of core collections for association studies: first, to study the global collection with molecular markers (data are frequently available because they are used for accession identification), and ideally, for phenotypic characteristics, at least those related to the trait/s of interest. The aim is to have enough richness to avoid having excessive redundancy. Second, to study redundancy to establish the size of the core subcollections, they should account for 90–99 % of the total richness. Third, to establish the size of the whole core collection, which will depend on several factors, including the genetic complexity of the trait of interest and the power required for the association study, but it should be, at least, three times the size of the core subcollections, to take advantage of the proposed approach. Finally, to build the core subcollections according to methods 2 or 3 proposed here, and merge the subcollections generated to render the whole core collection. It is convenient to study the possible stratification existing in the global collection to consider the possibility of creating core subcollections independently in each of the populations observed.

Conclusions

Association analysis is a strategy widely used in crop genetics at present, but it faces serious problems regarding the detection power and spurious associations due to diverse factors. One of them is the low frequency of certain classes of phenotypic traits in the collections used to this purpose. In this work, we propose two methods (2 and 3) to generate core collections in several sequential steps using M strategy, which could increase the representation of minority classes in the core collection and improve association results, reducing the rate of false negatives related to this problem. Method 2 is preferentially proposed because it is simpler than Method 3 and consists in the merge of several subcollections, each accounting for 90–99 % of richness and with the highest Shannon Index.

Acknowledgments This study was made possible with the funding from the GrapeGen project (joint venture between Genome Canada and Genoma España) and the AGL2010-15694 (MINECO) and AGL2014-59171-R (MINECO, FEDER). A.M. Vargas was funded by a predoctoral fellowship from Instituto Madrileño de Investigación y Desarrollo Rural, Agrario y Alimentario. Thanks to Rafael Torres-Pérez for helping with the structure analysis and to Joaquín Borrego, Carmen Fajardo, Carlos González Guillén, Mª Dolores Vélez, Silvia Hernáiz, Paz Fernández, Nuria Rodríguez Jiménez, and Concepción López Rivas for their technical assistance in the grapevine morphological descriptions.

Data archiving statement The phenotypic dataset for 1000 rice accessions was downloaded from http://www.genebank.go.kr/eng/PowerCore/ (Kim et al. 2007), and it is included in the supplementary material. The grapevine dataset is also included in the supplementary material, with a list of the accessions used. Additional information about the accessions of IMIDRA grapevine collection (ESP080) can be found in the website http://www.madrid.org/coleccionvidencin/.

Compliance with ethical standards

Conflict of interests The authors declare that they have no conflict of interests.

Research involving human participants and/or animals This article does not contain any studies with human participants or animals performed by any of the authors.

References

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y et al (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465:627–631. doi:10.1038/nature08800

Bacilieri R, Lacombe T, Le Cunff L, Vecchi-Staraz MD, Laucou V, Genna B, Péros J-P, This P, Boursiquot J-M (2013) Genetic structure in cultivated grapevines is linked to geography and human selection. BMC Plant Biol 13. doi: 10.1186/1471-2229-13-25

Bataillon TM, David JL, Schoen DJ (1996) Neutral genetic markers and conservation genetics: simulated germplasm collections. Genetics 144:409–417

Bordes J, Ravel C, Jaubertie JP, Duperrier B, Gardet O, Heumez E, Pissavy AL, Charmet G, Gouis JL, Balfourier F (2013) Genomic regions associated with the nitrogen limitation response revealed in a global wheat core collection. Theor Appl Genet 126:805–822

Brown AHD (1989) Core collections—a practical approach to genetic-resources management. Genome 31:818–824

Carpio DPD, Basnet RK, Vos RCHD, Maliepaard C, João M, Paulo BG (2011) Comparative methods for association studies: a case study on metabolite variation in a *Brassica rapa* core collection. PLoS One 6, e19624. doi:10.1371/journal.pone.0019624

Escribano P, Viruel MA, Hormaza JI (2008) Comparison of different methods to construct a core germplasm collection in woody perennial species with simple sequence repeats markers. A case study in cherimoya (*Annona cherimola*, Annonaceae) an underutilised subtropical fruit tree species. Ann Appl Biol. doi:10.1111/j.1744-7348.2008.00232.x

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 (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587

Fernandez L, Le Cunff L, Tello J, Lacombe T, Boursiquot JM, Fournier-Level A, Bravo G, Lalet S, Torregrosa L, This P, Martinez-Zapater JM (2014) Haplotype diversity of VvTFL1A gene and association with cluster traits in grapevine (*V. vinifera*). BMC Plant Biol 14. doi: 10.1186/s12870-014-0209-3

Franco J, Crossa J, Taba S, Shands H (2005) A sampling strategy for conserving genetic diversity when forming core subsets. Crop Sci 45:1035–1044. doi:10.2135/cropsci2004.0292

Franco J, Crossa J, Warburton ML, Taba S (2006) Sampling strategies for conserving maize diversity when forming core subsets using genetic markers. Crop Sci 46:854–864

Frankel OH (1984) Genetic perspectives of germplasm conservation. Genetic manipulation: impact on man and society. Cambridge University Press, Cambridge, pp 161–170

Gonzalez-Martinez SC, Huber D, Ersoz E, Davis JM, Neale DB (2008) Association genetics in *Pinus taeda* L. II. Carbon isotope discrimination. Heredity 101:19–26

Gouesnard B, Bataillon TM, Decoux G, Rozale C, Schoen DJ, David JL (2001) MSTRAT: an algorithm for building germ plasm core collections by maximizing allelic or phenotypic richness. J Hered 92:93–94

Hardy OJ, Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol Ecol Notes 2:618–620

Holbrook CC, Anderson WF (1995) Evaluation of a core collection to identify resistance to late leafspot in peanut. Crop Sci 35:1700–1702

Ibáñez J, Vargas AM, Palancar M, Borrego J, de Andrés MT (2009) Genetic relationships among table-grape varieties. Am J Enol Vitic 60:35–42

Khan M, Korban S (2012) Association mapping in forest trees and fruit crops. J Exp Bot 63:4045–4060. doi:10.1093/jxb/ers105

Kim K-W, Chung H-K, Cho G-T, Ma K-H, Chandrabalan D, Gwag J-G, Kim T-S, Cho E-G, Park Y-J (2007) PowerCore: a program applying the advanced M strategy with a heuristic search for establishing core sets. Bioinformatics 23:2155–2162

Kwon S-J, Brown AF, Hu J, McGee R, Watt C, Kisha T, Timmerman-Vaughan G, Grusak M, McPhee KE, Coyne CJ (2012) Genetic diversity, population structure and genome-wide marker-trait association analysis emphasizing seed nutrients of the USDA pea (*Pisum sativum* L.) core collection. Genes & Genomics 34:305–320

Le Cunff L, Fournier-Level A, Laucou V, Vezzuli S, Lacombe T, Adam-Blondon A, Boursiquot JM, This P (2008) Construction of nested genetic core collections to optimize the exploitation of natural diversity in *Vitis vinifera* L. subsp *sativa*. BMC Plant Biol 8. doi: 10.1186/1471-2229-8-31

Li X, Yan W, Agrama H, Jia L, Shen X, Jackson A, Moldenhauer K, Yeater K, McClung A, Wu D (2011) Mapping QTLs for improving grain yield using the USDA rice mini-core collection. Planta 234:347–361

Liu X, Huang M, Fan B, Buckler ES, Zhang Z (2016) Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS Genet 12(2), e1005767. doi:10.1371/journal.pgen.1005767

McKhann HI, Camilleri C, Bérard A, Bataillon TM, David JL, Reboud X, Le Corre V, Caloustian C, Gut IG, Brunel D (2004) Nested core collections maximizing genetic diversity in *Arabidopsis thaliana*. Plant J 38:193–202

Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang ZW, Costich DE, Buckler ES (2009) Association mapping: critical considerations shift from genotyping to experimental design. Plant Cell 21:2194–2202

Odong TL, Jansen J, van Eeuwijk FA, van Hintum TJL (2013) Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. Theor Appl Genet126. doi: 10.1007/s00122-012-1971-y

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

Ritland K (1996) Estimators for pairwise relatedness and individual inbreeding coefficients. Genet Res 67:175–185

Ronfort J, Bataillon T, Santoni S, Delalande M, David JL, Prosperi J-M (2006) Microsatellite diversity and abroad scale geographic structure in a model legume: building a set of nested core collection for studying naturally occurring variation in *Medicago truncatula*. BMC Plant Biol 6:28–40

Schoen DJ, Brown AHD (1993) Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. Proc Natl Acad Sci 90:10623–10627

Shin J, Lee C (2015) Statistical power for identifying nucleotide markers associated with quantitative traits in genome-wide association analysis using a mixed model. Genomics 105:1–4

Soto-Cerda BJ, Diederichsen A, Ragupathy R, Cloutier S (2013) Genetic characterization of a core collection of flax (Linum usitatissimum L.) suitable for association mapping studies and evidence of divergent selection between fiber and linseed types. BMC Plant Biol 13. doi: 10.1186/1471-2229-13-78

Soto-Cerda BJ, Duguid S, Booker H, Rowland G, Diederichsen A, Cloutier S (2014) Association mapping of seed quality traits using the Canadian flax (*Linum usitatissimum* L.) core collection. Theor Appl Genet 127:881–896. doi:10.1007/s00122-014-2264-4

Upadhyaya H, Wang Y, Sharma S, Singh S (2012) Association mapping of height and maturity across five environments using the sorghum mini core collection. Genome 55:471–479. doi:10.1139/g2012-034

Upadhyaya H, Wang Y, Gowda C, Sharma S (2013) Association mapping of maturity and plant height using SNP markers with the sorghum mini core collection. Theor Appl Genet 126:2003–2015. doi:10.1007/s00122-013-2113-x

van Hintum TJL, Brown AHD, Spillane C, Hodgkin T (2003) Colecciones núcleo de recursos fitogenéticos. Boletín Técnico del IPGRI 3

Vargas A, Fajardo M, Borrego J, de Andrés MT, Ibáñez J (2013a) Polymorphisms in VvPel associate with variation in berry texture and bunch size in the grapevine. Aust J Grape Wine Res 19:193–207. doi:10.1111/ajgw.12029

Vargas AM, Le Cunff L, This P, Ibáñez J, de Andrés MT (2013b) VvGAI1 polymorphisms associate with variation for berry traits in grapevine. Euphytica 191:85–98. doi:10.1007/s10681-013-0866-6

Wang ML, Sukumaran S, Barkley NA, Chen Z, Chen CY, Guo B, Pittman RN, Stalker HT, Holbrook CC, Pederson GA, Yu J (2011) Population structure and marker–trait association analysis of the US peanut (*Arachis hypogaea* L.) mini-core collection. Theor Appl Genet 123:1307–1317

Weir BS (2010) Statistical genetic issues for genome-wide association studies. Genome 53(11):869–875

Whitt S, Buckler E (2003) Using natural allelic diversity to evaluate gene function. Methods Mol Biol 236:123–140

Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42(4):355–360. doi:10.1038/ng.546

Zhao J, Artemyeva A, Del Carpio D, Basnet R, Zhang N, Gao J, Li F, Bucher J, Wang X, Visser R, Bonnema G (2010) Design of a *Brassica rapa* core collection for association mapping studies. Genome 53:884–898. doi:10.1139/G10-082

Zorić M, Dodig D, Kobiljski B, Quarrie S, Barnes J (2012) Population structure in a wheat core collection and genomic loci associated with yield under contrasting environments. Genetica 140:259–275