Statistical Techniques for Defining Reference Sets of Accessions and Microsatellite Markers

T. L. Odong,* J. van Heerwaarden, J. Jansen, T. J. L. van Hintum, and F. A. van Eeuwijk

ABSTRACT

Exploitation of the available genetic resources around the world requires information about the relationships and genetic diversity present among genebank collections. These relations can be established by defining for each crop a small but informative set of accessions, together with a small set of reliable molecular markers, that can be used as reference material. In this study, various strategies to arrive at small but informative reference sets are discussed. For selection of accessions, we proposed genetic distance optimization (GDOpt) method, which selects a subset of accessions that optimally represent the accessions not included in the core collection. The performance of GDOpt was compared with Core Hunter, an advanced stochastic local search algorithm for selecting core subsets. For the selection of molecular markers, we evaluated (i) the backward elimination (BE) method and (ii) methods based on principal component analysis (PCA). We examined the performance of the proposed methodologies using five real datasets. Relative to average distance between an accession and the nearest selected accession (representativeness), GDOpt outperformed Core Hunter. However, Core Hunter outperformed GDOpt with respect to allelic richness. The BE performed much better than other methods in selecting subsets of markers. Methods based on PCA showed that, for practical purposes, the inclusion of the first few (two or three) principal components (PCs) was often sufficient. To obtain robust and high-quality reference sets of accessions and markers we advise a combination of GDOpt (for accessions) and BE or methods based on PCA using a few PCs (for subsets of markers).

T.L. Odong, J. Jansen, and F.A. van Eeuwijk; Biometris, Wageningen Univ. and Research Centre, P.O. Box 100, 6700 AC Wageningen, the Netherlands; J. van Heerwaarden, Dep. of Plant Sciences, Univ. of California, Davis, CA 95616; T.J.L. van Hintum; Centre for Genetic Resources, The Netherlands (CGN), Wageningen, the Netherlands. Received 21 Feb. 2011. *Corresponding author (thomas.odong@wur.nl).

Abbreviations: BE, backward elimination; GCP, Generation Challenge Programme; GDOpt, genetic distance optimization; MAGIC, multiparent advanced generation intercross; PC, principal component; PCA, principal component analysis; PIC, polymorphism information content; PSA, proportion of shared alleles; QTL, quantitative trait loci; SNP, single nucleotide polymorphism; SSR, simple sequence repeat; WPCA, weighted principal component analysis.

PLANT GENETIC RESOURCES stored in genebanks offer great opportunities for improving and securing crop production, especially in marginal environments. Exploitation of the full potential of all available genetic resources around the world requires knowledge about the relationships relative to genetic diversity among genebank collections stored in different centers. The relations between genebank collections can be established by defining for each crop a small but informative set of accessions, together with a small set of reliable molecular markers, that can be used as reference material. Hereafter, the reference material will be referred to as "reference sets."

A reference set of a crop should be an adequate representation of the genetic diversity of that crop as stored in genebanks around the world. In that case, markers can be used to place new accessions in the spectrum of current accessions. The reference sets can also be used to connect different population genetic and quantitative genetic studies, including association studies.

Published in Crop Sci. 51:2401–2411 (2011).
doi: 10.2135/cropsci2011.02.0095
Published online 12 Aug. 2011.

Table 1. Summary information on the five data sets used in this study.

Crop	Number of accessions	Origins of accessions (no. of accessions in parentheses)	Number of SSR† markers
Coconut	1014	West Africa (32), North America (52), South Asia (62), Latin America (72), Central America and the Caribbean (109), East Africa (124), Southeast Asia (183), and the Pacific Islands (380).	30
Potato	233	Peru (91),Colombia (80), Bolivia (44), Ecuador (16), Argentina (1), and Chile (1).	50
Common bean	603	Peru (184), Mexico (178), Guatemala (6), Ecuador (135), Colombia (29), Brazil (22), and others (18 countries) (94).	36
Rice	1988	India (320), Bangladesh (210), China (167), Indonesia (166), Philippines (139), Liberia (137), Sri Lanka (124), Thailand (122), United States (99), Malaysia (97), Madagascar (87) Nigeria (80), and others (250).	37
Chickpea	3000	India (820), Iran (552), Syria (183), Turkey (160), Afghanistan (147), ICRISAT collections of mixed origin (138), Ethiopia (124), and others (876).	50

†SSR, simple sequence repeat.

To obtain reliable reference sets, large numbers of accessions have to be genotyped with markers. Under the auspices of the Generation Challenge Programme (GCP) (Generation Challenge Programme, 2008), large numbers of accessions of important agricultural crops were genotyped with 15 to 50 microsatellite markers. The GCP is a broad network of partners from international agricultural research institutes and national agricultural research programs collectively working to improve crop productivity in the developing world, especially environments prone to drought and having low soil fertility and high incidences of pests and diseases.

The general philosophy underlying the current study is that molecular markers, such as microsatellites, can be used to represent accessions as points in a multidimensional genetic space. A strategy for selecting accessions may consist of choosing accessions in such a way that the whole of the original genetic space is covered by a predefined number of accessions. With regard to molecular markers, the reference set should be able to approximate the full genetic space by preserving the distances between the accessions. It may be useful to identify clusters of accessions and use them as a basis for choosing accessions in a stratified way. In addition to statistical principles, molecular genetic requirements should be taken into account, especially the ease of generating markers and marker quality.

The concept of reference sets of accessions and markers is quite similar to the concept of forming core collections using marker information. The reference sets, unlike core collections, place emphasis on the selection of both accessions and molecular markers. In this case, the selected accessions are not linked to a specific genebank collection but taken from collections assembled from many centers. Brown (1995) referred to such a subset of accessions as synthetic core.

In this paper, various strategies to arrive at small but informative reference sets will be discussed. For selection of accessions, we propose a method based on optimization of the spacing of a fixed number of accessions within the genetic space; this method will be referred to as genetic distance optimization (GDOpt) method. To the best of our knowledge, no method currently exists for the selection of core collections that aims at obtaining a set of entries to maximize the representation of the accessions in the whole collection. Compared to GDOpt, most existing algorithms for selection of core collections (e.g., MSTRAT [Gouesnard et al., 2001], PowerCore [Kim et al., 2007], and Core Hunter [Thachuk et al., 2009]) pay more attention to the content of the core collections but tend to ignore the relationships between the selected entries and those not included in the core collection. The D-method (Franco et al., 2005) maximizes the representation of the groups with the assumption that the groups are known. The GDOpt aims specifically at the selection of core entries that optimally represent accessions not included in the core collection. For the selection of molecular markers, we examined (i) a backward elimination (BE) method and (ii) methods based on principal component analysis (PCA). Materials and Methods contains a description of the proposed methods and of five datasets used for illustration in this paper. In Results, the results of the application of the proposed methodologies to five datasets will be presented.

MATERIALS AND METHODS

Data

Coconut (Cocos nucifera L.)

The coconut data consist of 1014 accessions genotyped with 30 simple sequence repeat (SSR) markers. The accessions were collected from different regions of the world (see Table 1). Coconut is a diploid, mainly outcrossing species. Most of the accessions in this collection were described as tall; only 43 dwarf accessions, mainly from Southeast Asia, were present. Dwarf coconuts have a high degree of self-fertilization. More than half (19) of the 30 SSR markers used in this study have known positions on the linkage map; they are well spread across the genome.

Potato (Solanum Species)

The potato data consisted of 233 diploid accessions from four species [*S. ajanhuiri* Juz. & Bukasov (22 accessions), *S. tuberosum* L. subsp. *andigenum* (Juz. & Bukasov) Hawkes (syn. *S. goniocalyx* Juz. & Bukasov) (47 accessions), *S. tuberosum* L. subsp. *andigenum* (syn. *S. phureja* Juz. & Bukasov) (105 accessions), and *S. tuberosum* L. subsp. *andigenum* (syn. *S. stenotomum* Juz. & Bukasov) (59 accessions)] genotyped with 50 SSR markers (see Table 1). Potatoes are mainly outcrossing, with a substantial amount

of self-fertilization. The linkage group of 42 of the 50 SSR markers used in this study is currently known.

Common Bean (Phaseolus vulgaris L.)

Genotyped with 36 SSR markers, the common bean dataset consisted of 603 accessions with 296 being described as Andean and 307 as Mesoamerican types (see Table 1). Common bean is a self-pollinating diploid species. Twenty-nine of the 36 SSR markers used in study belong to known linkage groups.

Rice (Oryza sativa L.)

The rice dataset consisted of 1998 accessions genotyped with 37 markers (see Table 1). Rice is a self-pollinating diploid species. The linkage map positions of all 37 SSR markers used in study are known.

Chickpea (Cicer arietinum L.)

The chickpea data consisted of 3000 accessions genotyped with 50 SSR markers. The accessions originated from more than 60 countries (mainly from the Middle East and other parts of Asia), with germplasm collections maintained at two international centers (ICRISAT in India and ICARDA in Syria) and at several national genebanks (see Table 1). Chickpea is a self-pollinating diploid species. Thirty-two of the 50 SSR markers used in study have known linkage groups but the positions of the markers on the linkage map were not available.

Strategies for Selecting Representative Accessions

A number of strategies for selecting subsets from large collections of accessions (with special reference to the forming of core collections) have been proposed: MSTRAT (Gouesnard et al., 2001), genetic distance sampling (Jansen and van Hintum, 2007), PowerCore (Kim et al., 2007) and Core Hunter (Thachuk et al., 2009). With the exception of genetic distance sampling, all methods mentioned above apply the M-strategy (Schoen and Brown, 1993); the M-strategy aims at maximizing the number of observed alleles of the markers in the subset of selected accessions. In genetic distance sampling, accessions are selected in such a way that selected accessions are always a predefined distance (selection radius) away from each other. This ensures that no duplicates or similar accessions are selected. A disadvantage of the M-strategy is that it is likely to select nonrepresentative accessions ("outliers"). None of the above methods was developed to select accessions to serve as representatives around which the other accessions can be positioned. In this paper, we propose GDOpt for selecting representative accessions.

Genetic Distance Optimization

The aim of GDOpt is to select a fixed number (say K) of representative accessions. It is a form of K-medoids clustering (Kaufman and Rousseeuw, 1990), in which one accession in each of K clusters acts as center of the cluster. Clusters are formed by minimizing the total distance of all accessions to the nearest of the K accessions designated as cluster centers. The current algorithm utilizes simulated annealing (Kirkpatrick et al., 1983). To obtain a good starting point, the initial configuration of cluster centers is provided by a modified version of genetic distance sampling (Jansen and van Hintum, 2007). Genetic distance sampling was modified to select a fixed number of accessions by adjusting the selection radius until the number of accessions selected by genetic distance sampling was equal to or greater than the required size of the reference set. If the number of accessions selected by genetic distance sampling is greater than the intended size of the reference set, random sampling is used to delete the extras. Eventually, the algorithm will be made available as a procedure in the Biometris Genstat Library (http://www.biometris.wur.nl/UK [verified 22 July 2011]) but at the moment it is available only on request from the authors.

Comparison with Core Hunter

In this paper, the results obtained with GDOpt are compared with those obtained with Core Hunter (Thachuk et al., 2009). Core Hunter was selected because the authors have demonstrated its superiority over other existing methods of core selection. In Core Hunter, the weights attached to two optimization criteria (modified Rogers distance and Shannon diversity index) were varied. By assigning all the weight to the modified Rogers distance, Core Hunter maximizes the average genetic distance between selected accessions, whereas by assigning all the weight to Shannon diversity index, it maximizes the number of alleles in the selected accessions. The comparison was based on two criteria: (i) the distance between accessions and the nearest entry in the reference set (representativeness) and (ii) the proportion of alleles captured in a subset of a specified sample size selected by each method. This comparison was done to show that forming core collections with the intention to maximize either allelic richness or distances between entries (e.g., using Core Hunter settings in this study) compromises the ability to represent the contents of the whole collection.

The results from GDOpt and Core Hunter were also compared with those from simple random sampling (for real data) and stratified random sampling for simulated data. The details on the simulation and analysis of the simulated datasets are presented as Supplemental file Appendix 1.

Selecting Subsets of Molecular Markers

Criterion

In the current context, the criterion used for comparing different methods of selecting subsets of molecular markers is based on the preservation of genetic distances between accessions. The key assumption is that by preserving genetic distances between accessions, population structure (if present) will be preserved. The criterion applied in all cases is the correlation between genetic distances between accessions based on a subset of molecular markers and genetic distances based on all available markers.

Polymorphism Information Content

The polymorphism information content (PIC) (Botstein et al., 1980) depends on the number and frequencies of alleles. According to this criterion, a marker with many alleles with small frequencies is more informative than a marker with two alleles with equal frequencies. The PIC does not take into account the dependencies between markers. Because it is one of the most frequently used criteria for selecting sets of molecular markers, the performance of other methods will be compared with that based on PIC.

Methods Based on Principal Component Analysis

These methods use the dimension reduction ability of PCA to identify a subset of molecular markers that should be retained to achieve minimum loss of information. Recently, the use of PCA for selecting subsets of molecular markers (especially single nucleotide polymorphisms [SNPs]) has been discussed by Paschou et al. (2007) and Zhang et al. (2009).

Molecular markers are selected based on the weighted sum of squared loadings on all principal components (PCs) designated as important, using the corresponding eigenvalues as weights. The method will be referred to as weighted PCA (WPCA). The steps are as follows: (i) perform PCA on the accession-by-marker matrix, (ii) decide on the number of PCs to be designated as important, (iii) calculate the weighted sum of squares of the loadings of each marker on the PCs designated as important, and (iv) rank the markers in descending order based on their weighted sums of squared loadings. The molecular markers are then included in the subset based on their ranks. For WPCA, we compared (i) ranking based on the first PC (WPCA1), (ii) ranking based on the first two PCs (WPCA2), (iii) ranking based on the first three PCs (WPCA3), and (iv) ranking based on the first 20 PCs (WPCA20) when selecting a subset of markers.

Patterson et al. (2006) discussed the use of the Tracy–Widom distribution for determining the number of significant PCs for SNP data. This is done by comparing standardized eigenvalues with the Tracy–Widom distribution. If n differentiated groups of genotypes are present in the data, one expects to find $k = n - 1$ significant eigenvalues. Van Heerwaarden, Odong, and van Eeuwijk (unpublished data, 2011) suggested a modification of the above method for SSR markers. However, in practice it has become standard to designate the first two or three PCs as important and discard the rest without performing any statistical test. Formal testing usually leads to many statistically significant PCs.

Application of PCA to SSR data requires special attention. The SSR markers were first recoded as 0, 1, or 2 based on the number of copies of the allele with frequency closest to 0.5. The advantage of treating SSR markers in this way lies in its simplicity. We expect the loss of information associated with coding SSR markers in this way to be small in most cases. The SSR marker data were recoded as described above to reduce the information from each SSR marker into a single column, which can then be easily related to PCs.

Backward Elimination

This method is similar to backward elimination method used for variable selection in multiple regression. It uses the correlation between the genetic distances (between accessions) based on all molecular markers and the genetic distances based on a subset of markers as the criterion for deleting markers. In a stepwise approach, at each step, the molecular marker whose exclusion leads to the smallest reduction in correlation between the two sets of distances is removed until a specified level of correlation or a desired number of molecular markers is reached.

The BE method can be summarized as follows:

Step 1: Calculate the distances between accessions using all the molecular markers. Let D_0 be the matrix of those distances ($D_0 = d_{ij}$, where d_{ij} is the distance between accession i and j).

Step 2: For each of the m markers, calculate the distances between accessions by leaving out one marker at a time. Let D_{-e} ($e = 1,2,\ldots,m$) be the matrix of distances between accessions constructed with marker e left out ($D_{-e} = d_{ije}$, where d_{ije} is the distance between accessions i and j calculated when marker e is left out). Denote r_{-e} as the correlation between D_0 and D_{-e}.

$$r_{-e} = [\sum_{i<j}^{n}(d_{ij} - \bar{d})(d_{ije} - \bar{d}_{ije})] / [\sum_{i<j}(d_{ij} - \bar{d})^2 \sum_{i<j}(d_{ije} - \bar{d}_{ije})^2].$$

Step 3: Select the marker with the largest r_{-e} value, eliminate it (the marker) from the dataset, and repeat Step 2 with the remaining markers. Each time, the maximum value of r_{-e} is recorded.

Step 4: Repeat steps 2 and 3 until either the maximum value of r_{-e} reaches the stopping value set or until the desired number of markers is achieved.

Similarity Measures

In this paper, we used genetic distances (D) based on the proportion of shared alleles (PSA) applied to the original SSR marker data and the recoded data, where $D = 1 - \text{PSA}$, and

$$\text{PSA} = \left[\sum_{n=1}^{M}\sum_{a=1}^{A_m}\min(f_{1ma}, f_{2ma})\right] / M,$$

where f_{1ma} and f_{2ma} are the frequencies of allele a ($a = 1,2,\ldots,A_m$) for molecular marker m ($m = 1,2,\ldots,M$) in individuals 1 and 2, respectively. For more information on the PSA as similarity measure, see Bowcock et al. (1994), Chakraborty and Jin (1994), and Chang et al. (2009).

Other Important Aspects of Selecting Subsets of Molecular Markers

In addition to the statistical criteria used for selecting molecular markers, a number of important issues should also be examined. The nonstatistical issues of importance in marker selection are quality relative to clarity and repeatability of banding pattern, ease of automation of allele calling, and genome coverage and linkage between markers. The markers selected should be of high quality with highly reproducible alleles.

RESULTS

Selection of Accessions

General Results

In the following, representativeness is measured as average distance between each accession to the nearest selected entry in the subset of accessions (Table 2). The GDOpt produces subsets of accessions that are much more representative compared with Core Hunter. In all the crops, the average distance from accessions to its nearest entry in the subset of accessions is smaller for GDOpt compared with all settings of Core Hunter. Random sampling also performed much better than all the different settings of Core Hunter relative to representativeness of the whole collection.

With regard to the total number of alleles captured by subsets of 15 selected accessions (Table 3), all parameter settings of Core Hunter performed better than GDOpt. However, major

Table 2. Average distances between accessions and their nearest entry in the selected subset of accessions obtained using genetic distance optimization (GDOpt), random sampling, and Core Hunter (CH) with five (CH1–CH5) different parameter settings in terms of modified Rogers distance (MR) and Shannon diversity index (SH). Random sampling values were obtained from 100 samples.

Method	Crop				
	Coconut	Potato	Common bean	Rice	Chickpea
GDOpt	0.389	0.216	0.359	0.472	0.646
Random sampling	0.463	0.274	0.443	0.548	0.729
CH1 (MR = 1.0; SH = 0.0)†	0.490	0.307	0.467	0.547	0.760
CH2 (MR = 0.7; SH = 0.3)	0.522	0.325	0.476	0.551	0.775
CH3 (MR = 0.5; SH = 0.5)	0.531	0.327	0.478	0.542	0.760
CH4 (MR = 0.3; SH = 0.7)	0.527	0.326	0.474	0.534	0.748
CH5 (MR = 0.0; SH = 1.0)	0.521	0.321	0.483	0.537	0.766

†The values in the parentheses show the different weights given to modified Rogers distance (MR) and Shannon diversity index (SH) used when selecting a subset of accessions using Core Hunter.

differences were found in the retention of alleles with different frequencies (see Fig. 1). For ease of interpretation, we have classified alleles into three categories based on their frequencies (p): (i) common alleles ($p \geq 0.05$), (ii) rare alleles ($0.005 \leq p < 0.05$), and (iii) very rare alleles ($p < 0.005$). The proportion of common alleles captured by GDOpt and different settings of Core Hunter were comparable. For all five crops, subsets of 15 accessions selected using GDOpt performed well in capturing common alleles. With the exception of chickpea, subsets selected via GDOpt captured more than 85% of all common alleles. In potato and common bean, subsets of accessions obtained by GDOpt showed a higher frequency of common alleles compared with subsets of accessions obtained by the different settings of Core Hunter. Core Hunter performed much better than GDOpt in capturing rare and very rare alleles. However, with simulated data, GDOpt performed better than Core Hunter with all the weights given to modified Roger's distance relative to proportion of captured alleles (see Supplemental file Appendix 1).

Selection of Markers

General Results

In the following, the preservation of pairwise distances between accessions by a subset of SSR markers is measured by the correlation between the distances based on the subset of SSR markers and the distances based on the whole set of SSR markers (Table 4).

Across all five crops, BE performed much better than all other methods in selecting a subset of molecular markers in preserving the pairwise distances between accessions. The selection based on PIC performed very poorly in datasets with very many alleles (common bean and chickpea). The method based on WPCA using many PCs (WPCA20) usually produced worse results compared with when one, two, or three PCs (WPCA1, WPCA2, or WPCA3) were used. The differences in performance between the methods became more pronounced when selecting small subsets (<10) of SSR markers (results not shown).

The number of SSR markers required to achieve a specified minimum correlation depended on whether the SSR markers are recoded or not (Table 5). For all five crops, fewer markers were required to achieve a specified correlation when the PSA was calculated from the original SSR data instead of recoded data. The differences can be attributed to the loss of information associated with recoding SSR markers and this loss of information appears to be large for SSR markers with high PIC values.

Evaluation of subsets of five SSR markers indicated that BE and WPCA-based methods tended to select markers whose major alleles had frequencies close to 0.5 (Table 6). These SSR markers separated major groups of accessions.

Table 3. Numbers of alleles in the whole datasets and proportions of alleles in subsets of 15 accessions obtained using genetic distance optimization (GDOpt) , random sampling, and Core Hunter (CH) with five (CH1–CH5) different parameter settings in terms of modified Rogers distance (MR) and Shannon diversity index (SH). Random sample values were obtained from 10 samples.

Method	Crop				
	Coconut	Potato	Common bean	Rice	Chickpea
Whole dataset	469	367	1089	566	1605
GDOpt	0.422	0.635	0.255	0.339	0.318
Random sampling	0.430	0.554	0.254	0.344	0.264
CH1 (MR = 1.0; SH = 0.0)†	0.388	0.700	0.298	0.426	0.318
CH2 (MR = 0.7; SH = 0.3)	0.527	0.796	0.332	0.459	0.338
CH3 (MR = 0.5; SH = 0.5)	0.563	0.820	0.341	0.466	0.336
CH4 (MR = 0.3; SH = 0.7)	0.569	0.837	0.346	0.463	0.333
CH5 (MR = 0.0; SH = 1.0)	0.569	0.839	0.350	0.482	0.343

†The values in the parentheses show the different weights given to modified Rogers distance (MR) and Shannon diversity index (SH) used when selecting a subset of accessions using Core Hunter.

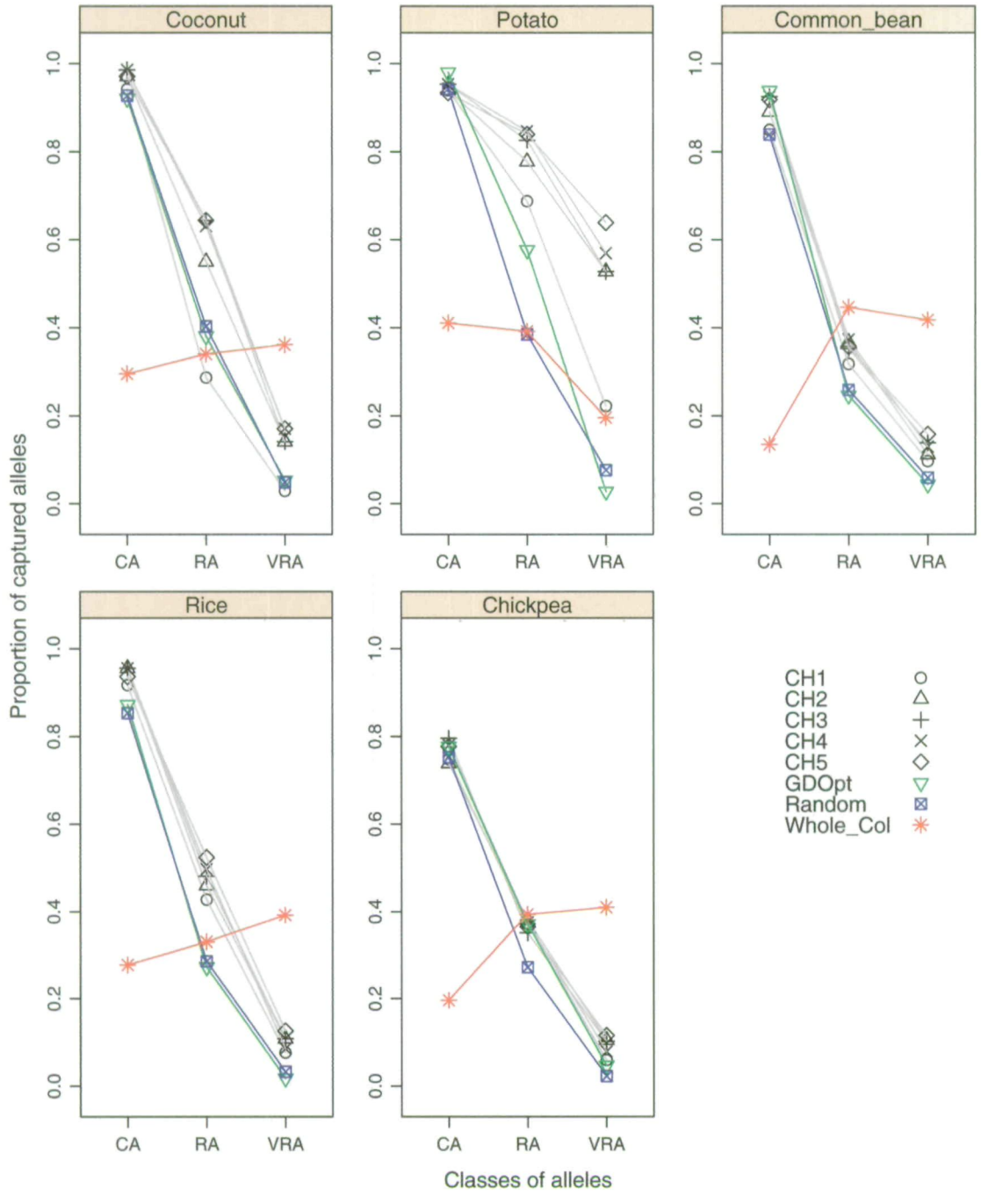

Figure 1. Proportions of alleles in different classes in the whole dataset (Whole_Col) and in subsets of 15 accessions obtained using genetic distance optimization (GDOpt), random sampling, and different parameter settings for Core Hunter (CH1-RD(1)SH(0); CH2-RD(0.7)SH(0.3); CH3-RD(0.5)SH(0.5); CH4-RD(0.3)SH(0.7); CH5-RD(0)SH(1)). The parameter settings refer to weights assigned to modified Rogers distance (RD) and Shannon diversity index (SH). Classes are based on the frequencies of the alleles in whole collection (common alleles [CA] [$p \geq 0.05$], rare alleles [RA] [$0.005 \leq p < 0.05$], and very rare alleles [VRA] [$p < 0.005$])

The PIC criterion favored SSR markers with very many alleles. These SSR markers differentiated between individual accessions or small groups of accessions and thus played a minimal role in separating major groups.

Crop-Specific Results

Coconut

The best method for selection of a subset of markers was BE, followed by WPCA1 and WPCA2 (Fig. 2). For example, when distances were based on PSA, using the original SSR marker data we only required seven out 30 markers using BE, compared with 14 using PIC criterion to achieve a correlation of 0.85. The results obtained for WPCA3 and WPCA20 were quite similar to random sampling of marker subsets but better than those for PIC. A similar pattern in the number of molecular markers required to achieve a correlation of 0.85 was observed when distances were calculated using the recoded data, except that the numbers of required markers were much higher.

Table 4. Correlation of pairwise distances between accessions for a subset of five markers versus all the markers with distance based on the proportion of shared alleles (PSA).

	Crop				
Method†	Coconut	Potato	Common bean	Rice	Chickpea
BE	0.813	0.826	0.902	0.706	0.661
WPCA1	0.775	0.718	0.864	0.698	0.640
WPCA2	0.766	0.772	0.722	0.533	0.624
WPCA3	0.653	0.651	0.719	0.407	0.624
WPCA20	0.669	0.607	0.617	0.361	0.535
PIC	0.603	0.663	0.527	0.607	0.347

†BE, backward elimination; WPCA1, WPCA2, WPCA3, and WPCA20, weighted principal component analysis using the first 1, 2, 3, and 20 principal components, respectively; PIC, polymorphic information content.

Potato

Backward elimination outperformed all other methods in selecting a subset of molecular markers (Fig. 2). When pairwise distances between accessions were calculated using PSA based on the original SSR marker data, we only needed six out of 50 markers to achieve a correlation of 0.85, which is less than half the number required by other methods. For the number of molecular markers needed to achieve a correlation of 0.85 (with the exception of WPCA2), the performances of the other methods were quite similar. Only BE, WPCA1, and WPCA2 performed better than random selection.

Common Bean

For common bean, a much greater difference in the performance of BE, compared with the other methods (except WPCA1), was found than for the other crops, especially in subsets of markers of small size (Fig. 2). The PIC performed very poorly in this dataset. The BE and WPCA1 required only two out 33 of SSR markers to achieve a correlation of 0.85 compared with 13 markers for PIC. The performance of WPCA20 was quite similar to that of PIC.

Rice

The BE performed better than the other methods, except WPCA1 (Fig. 2). The performances of BE, WPCA1 and WPCA2 were very similar for subsets of markers with sizes greater than 10. For subsets of markers of sizes less than 10, random selection of markers performed much better than WPCA3 and WPCA20. With the exception of BE and WPCA1, the method based on PIC performed better than other methods for subsets of size less than five. When correlation was based on recoded SSR data, WPCA20 and PIC required the same number of markers (22) to achieve a correlation of 0.85.

Chickpea

Although BE method performed better than all the other methods, the differences in performance were not prominent, especially with WPCA-based methods (Fig. 2). The selection based on PIC performed very poorly compared with the other methods. The PIC required 40 out of 50 markers to achieve a correlation of 0.85. In this case, randomly selecting a subset of SSR markers produced much better results than PIC.

Table 5. Numbers of selected markers required to achieve a minimum correlation of 0.85 between distances between accessions based on markers selected using different methods and on distances between accessions based on all markers. The numbers in the parenthesis is obtained when the distances between accessions were based on SSR data recorded as 0, 1, or 2.

	Crop				
Method†	Coconut	Potato	Common bean	Rice	Chickpea
BE	7 (12)	6 (13)	2 (3)	13 (18)	16 (23)
WPCA1	9 (16)	16 (23)	2 (2)	13 (18)	18 (26)
WPCA2	9 (14)	11 (24)	7 (11)	13 (18)	17 (26)
WPCA3	11 (15)	15 (24)	9 (12)	15 (19)	17 (24)
WPCA20	11 (16)	15 (28)	11 (14)	20 (22)	21 (25)
PIC	14 (20)	18 (31)	13 (17)	18 (22)	40 (40)

†BE, backward elimination; WPCA1, WPCA2, WPCA3, and WPCA20, weighted principal component analysis using the first 1, 2, 3, and 20 principal components, respectively; PIC, polymorphic information content.

DISCUSSION AND CONCLUSIONS

Understanding the current status of genetic diversity and finding links between genetic resources stored in different institutions are essential for a successful worldwide exploitation of genetic resources for crop improvement. The concept of reference sets of accessions and markers provides an efficient way to relate new materials to existing ones and set up different crop-specific study panels that can be used by plant breeders worldwide, with just a few representative accessions and a few molecular markers covering the genetic diversity in each crop. For example, selected accessions can be used for creating the so-called MAGIC (multiparent advanced generation intercross) population, which can be used for quantitative trait loci (QTL) analysis. Kover et al. (2009) demonstrated the utility of MAGIC population in improving the precision of QTL mapping.

In this study, representative accessions were selected using GDOpt, which aims at optimizing the spacing of a fixed number of accessions within the genetic space defined by all available markers. By performing selection and clustering of accessions simultaneously, this method can avoid the tedious process of determining population structure of the collection. Determination of population structure is quite challenging, especially in the case of germplasm collections where most often no clearly defined groups exist (Odong et al., 2011). In highly diverse collections, it may only be possible to isolate subsets of closely related individuals rather than obtaining large homogenous groups (Hamblin et al., 2007). It is from these closely related individuals that GDOpt selects a representative. Results from simulations have shown that if groups are known, stratified sampling does give improvement over simple random

Table 6. Average frequencies of major alleles in a subset of five simple sequence repeat (SSR) markers selected by different methods.

	Crop				
Method†	Coconut	Potato	Common bean	Rice	Chickpea
BE	0.501	0.511	0.484	0.311	0.420
WPCA1	0.515	0.566	0.515	0.459	0.435
WPCA2	0.562	0.541	0.320	0.411	0.461
WPCA3	0.388	0.482	0.455	0.440	0.461
WPCA20	0.367	0.428	0.386	0.414	0.209
PIC	0.241	0.357	0.122	0.201	0.070

†BE, backward elimination; WPCA1, WPCA2, WPCA3, and WPCA20, weighted principal component analysis using the first 1, 2, 3, and 20 principal components, respectively; PIC, polymorphic information content.

sampling, but its performance is still worse than that of GDOpt (Supplemental file Appendix 1). However, in situations where distinct groups of accessions exist (e.g., the Andean and Mesoamerican types of common beans), the selection can be performed separately for each group. Most methods that aim at optimizing either allelic richness or maximum genetic distances between selected accessions are quite capable of covering the full range of genetic diversity, including extremes, but may not produce representative subsets of accessions. For example, by simply selecting extremes, it would be possible to produce a subset with maximum genetic distances between accessions or maximum number of alleles although the selected accessions are not fully representative of the whole collection. Moreover, according to Zhang et al. (2010), the majority of very rare alleles would not contribute to the genetic diversity needed to develop elite cultivars and therefore their inclusion in the core collection may not be worthwhile. Some scientists (Allard, 1992; Frankel et al., 1995) have argued that less frequent alleles only occasionally affect quality or other traits and are generally unlikely to be of future use. In a situation where a representative subset is required, GDOpt has great advantages over all other methods, as shown in this study.

One of the key challenges in selecting representative sets of accessions based on distances between accessions is the effect of (random) errors in the data. In general, (random) errors will inflate dissimilarities between individuals, with smaller dissimilarities being relatively more inflated than larger ones. The inflation of dissimilarities consequently results in an overall greater dispersion of accessions in the genetic space, making it more difficult to obtain representative sets of accessions. The use of SSR markers with very many alleles (and consequently high PIC values) aggravates this problem. It is thus clear that if we are interested in a stable relationship between accessions, then the distances obtained from all the available markers and/or all alleles may be unsuitable. Markers with very high PIC (or very many alleles), in addition to inflating the distances between accessions, are likely to provide inconsistent relationships because of the fact that some of the alleles are as a result of misreading bands and are not repeatable. A much more stable relationship (distance) between accessions can be obtained by discarding some markers. Our results show that for all the five crops, 10 or more markers can be discarded without much distortion of pairwise distances between accessions. Another alternative for obtaining a stable relationship between accessions or group of accessions would be to calculate distances using important PCs, but additional studies are needed.

For the selection of subsets of molecular markers, we have shown that if one is interested in selecting a subset that preserves pairwise distances between accessions, BE provides the best option. The BE tends to remove markers with very many alleles and lots of missing values because they tend to contribute less to pairwise distances between accessions. The first markers included in the subset using BE mainly separate the major groups present in the data but could have the weakness of not differentiating well between accessions within groups. For example, for the common bean data, only two markers are required to achieve a correlation of 0.85 and those two markers separate Mesoamerican and Andean types quite well. A similar situation was observed for coconut, where the first five markers separated accessions associated with the Pacific Ocean from those associated with Indian and Atlantic Oceans. Simulations (results not shown) indicated that the correlation between pairwise distances between accessions based on a subset of markers and distances based on the entire set of markers depended on the level of group structure in the data. The stronger the group structure, the fewer the number of markers required to preserve the pairwise distances between accessions. The performance of BE could be improved by performing marker selections in two steps; that is, first perform BE based on the whole dataset and subsequently perform it within the major groups. For rice and chickpea, the difference in performance between BE and other methods was smaller compared to common bean, coconut. and potatoes. This could be attributed to the nature of group structure present in these datasets. Both multidimensional scaling and cluster analysis showed the presence of strong structure that was consistent with passport data in the three datasets (common bean, coconut, and potato), which indicated a large difference between BE and other methods of selection of subset of markers compared to rice and chickpea.

The performances of the PCA-based methods were quite good and in some cases comparable with that of BE. Our study revealed one interesting aspect about the number of important PCs to be included in the selection process. In all our datasets, the first few (1 to 3) PCs appeared to be sufficient. For most datasets, the eigenvalues revealed a big difference between the first two or three PCs compared with the rest; which made the contribution of the later PCs of minor importance. The practice of determining the number of important PCs through rigorous statistical testing most often leads to inclusion of too many PCs,

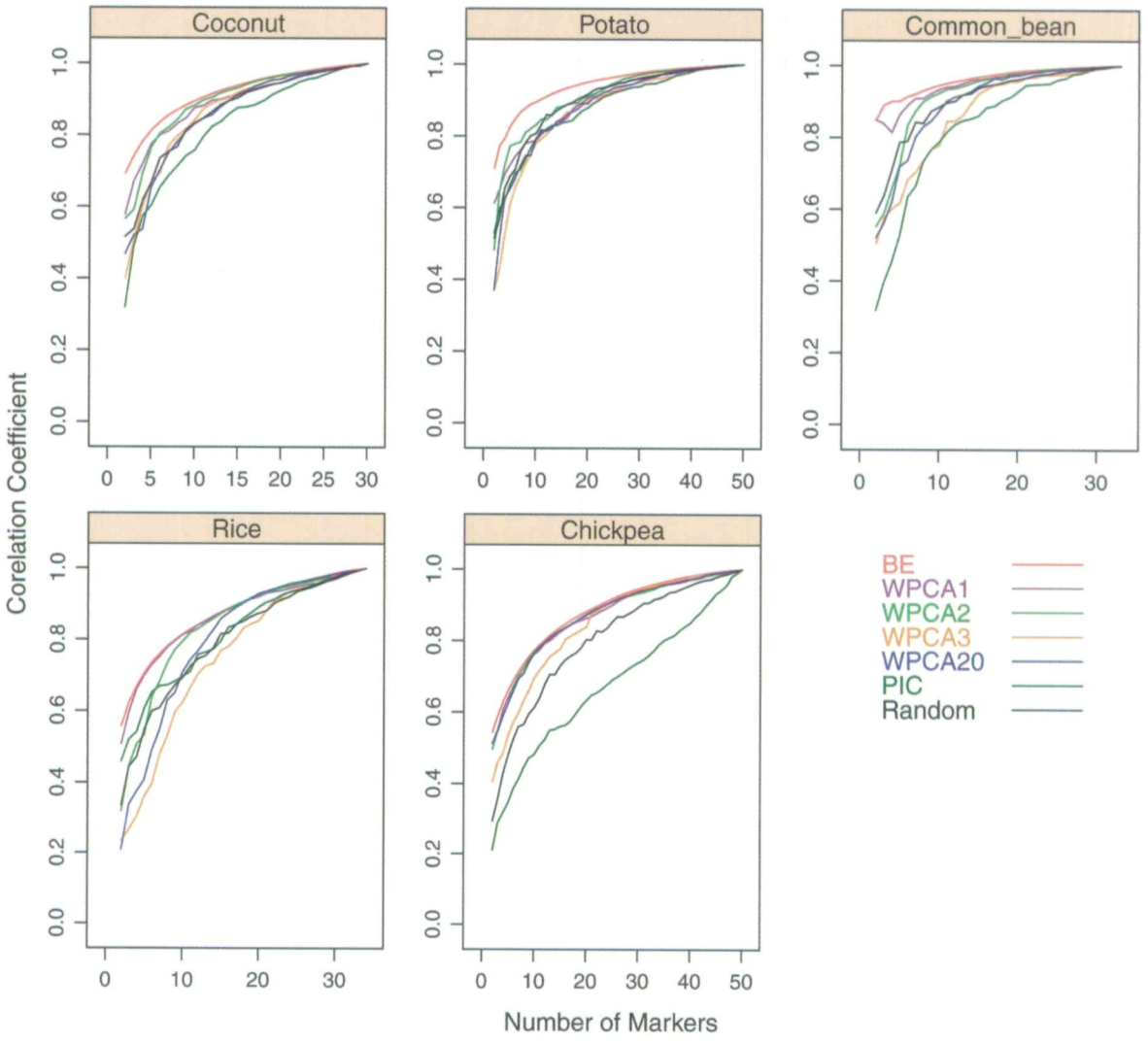

Figure 2. Correlation between distances constructed using a subset of simple sequence repeat (SSR) markers versus all the SSR markers for the different selection methods with distances based on the proportion of shared alleles (PSA). BE, backward elimination; WPCA1, WPCA2, WPCA3, and WPCA20, weighted principal component analysis using the first 1, 2, 3, and 20 principal components, respectively; PIC, polymorphism information content.

which in turn introduces noise. A recent study by Lee et al. (2009) noted a negative effect of including all significant PCs when performing distance-based cluster analysis.

Subsets of markers selected using PIC performed very poorly in preserving pairwise distances between the accessions, especially with common bean and chickpea. The poor performance with common bean and chickpea could be attributed to the poor quality of the data. Both datasets contained many markers with a very large number of alleles with more than 50% of the alleles having frequencies of less than 0.01 (see Supplemental file Appendix 2 for diversity statistics of the SSR markers used in this study). In both crop species, the average frequency of major alleles for the five SSR markers with the highest PIC is much smaller compared with subsets formed by BE and PCA-based methods. A large of number of alleles with frequencies of less than 0.01 are because of poor binning of alleles. The presence of error (random) in the data was thus more likely to affect selection of markers based on PIC compared with BE and WPCA-based methods. The BE and WPCA-based methods (especially WPCA1, WPCA2, and WPCA3) were more robust for detecting errors because those methods only picked out the key features of the data. Although PIC is the most common criterion used for selection of molecular markers, we have shown in this study that it performed poorly with respect to preserving major relationships between groups of individuals. Because PIC measures genetic diversity within a population, its poor performance with respect to identifying major features in the data is not surprising.

When SSR markers were recoded, the difference in performance between the different methods was smaller compared with the results obtained using the original SSR marker scores. This may be because of a loss of information;

forcing alleles into just two categories (allele with frequency closest to 0.5 versus others) tends to smooth out differences between accessions. It is clear from literature that one needs more biallelic markers to achieve the same level of genetic distance accuracy as a set of multiallelic markers, such as microsatellites (see Laval et al., 2002). As noted from the results in this study, recoding affected markers with a high PIC much more than other markers. The correlation between distances between the accessions based on the original SSR markers and distances based on recoded SSR markers indicated some loss of information. The correlations for chickpea, coconut, rice, common bean, and potatoes were 0.42, 0.69, 0.71, 0.82, and 0.88, respectively. The low correlation for chickpea (0.42) is an indication that recoding SSR data can sometimes lead to a substantial loss of information, and therefore it should be applied cautiously. Other methods, such as performing PCA on allele frequencies from each SSR marker separately and later combining the information across all markers, can be explored.

One of the key advantages of BE and PCA-based methods is that the selected molecular markers are likely to be independent. For PIC, unless sets of markers on which selection is done are known to be independent, there is no guarantee that the selected markers will be independent. For the datasets used in this study, several of the markers provided were on different linkage groups and those for which the positions on the chromosomes were given showed wide spacing between the markers (independence).

It is clear from our study that by using both BE and PCA-based methods, several good subsets of markers can be obtained. Other (quality) aspects of the chosen molecular markers (e.g., the possibilities for multiplexing) can be used to identify the most appropriate set. In the same way, alternative sets of accessions also exist and suitable accessions can be selected to replace less desirable ones. For example, accessions with missing values or those known to have propagation problems can be replaced. Discussion with genebank curators, crop specialists, and laboratory technicians can provide information that can be used as a basis of determining which of the selected accessions and molecular markers should be retained or dropped. The use of multivariate statistical techniques, such as multidimensional scaling, can assist in visualizing the selected accession in the space defined by the selected subset of markers.

In summary, for the selection of subsets of both accessions and markers, several methods exist, each with their own advantages and disadvantages; that is, there is no perfect core collection suitable for all purposes. Although GDOpt performs very well with respect to representativeness of nonselected accessions, its performance with respect to maximizing genetic diversity parameters, such as allelic richness or distances between selected accessions, is slightly compromised—that is, there is a trade-off. Methods such as MSTRAT (Gouesnard et al., 2001), PowerCore (Kim et al., 2007), and Core Hunter (Thachuk et al., 2009) should be used when the interest is in selecting subsets of accessions by maximizing diversity parameters, such as allelic richness or distance between entries in the core collection. For the selection of subsets of molecular markers, both BE and methods based on the first few (two or three) PCs gave rise to subsets of markers that preserved the major structure in the data but may have performed poorly for discriminating between individuals within the groups compared with markers with a high PIC.

Supplemental Information Available

Supplemental material is available free of charge at http://www.crops.org/publications/cs.

Acknowledgments

We thank the technical editor and two anonymous reviewers for their critical review of the manuscript. This work was supported by the Generation Challenge Programme under GCP subprogram I – Crop Genetic Diversity. We thank various people who participated in generating the data used in this study, especially Carmen de Vicente, Patricia Lebrun-Turquay (PI – coconut), Matthew Blair (PI – common bean), Marc Ghislain (PI – potato), D. Hoisington, H.D. Upadhyaya, and R. Varshney (PIs – chickpea), and Kenneth McNally, Claudio Brondani, Claire Billot, Mathias Lorieux, and Adam Famoso (PIs – rice).

References

Allard, R.W. 1992. Predictive methods for germplasm identification. p. 119–146. *In* H.T. Stalker and J.P. Murphy (ed.) Plant breeding in the 1990's. CAB International, Wallingford, UK.

Botstein, D., R.L. White, M. Skolnick, and R.W. Davis. 1980. Construction of a genetic-linkage map in human using restriction fragment length polymorphisms. Am. J. Hum. Genet. 32:314–331.

Bowcock, A.M., A. Ruizlinares, J. Tomfohrde, E. Minch, J.R. Kidd, and L.L. Cavallisforza. 1994. High-resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455–457. doi:10.1038/368455a0

Brown, A.H.D. 1995. The core collection at the crossroads. p. 77–92. *In* T. Hodgkin, A.H.D. Brown, Th.J.L. van Hintum, and E.A.V. Morales (ed.) Core collections of plant genetic resources. John Wiley & Sons, Chichester, UK.

Chakraborty, R., and L. Jin. 1994. Determination of relatedness between individuals using DNA-fingerprinting. Hum. Biol. 65:875–895.

Chang, W.H., H.P. Chu, Y.N. Jiang, S.H. Li, Y. Wang, C.H. Chen, K.J. Chen, C.Y. Lin, and Y.T. Ju. 2009. Genetic variation and phylogenetics of Lanyu and exotic pig breeds in Taiwan analyzed by nineteen microsatellite markers. J. Anim. Sci. 87:1–8. doi:10.2527/jas.2007-0562

Frankel, O.H., A.H.D. Brown, and J.J. Burdon. 1995. The conservation of plant biodiversity. Cambridge Univ. Press, UK.

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

Generation Challenge Programme. 2008. Generation Challenge

Programme Central registry. Available at http://www.generationcp.org (accessed Dec. 2008, verified 27 July 2011). Generation Challenge Programme, Texcoco, Mexico.

Gouesnard, B., T.M. Bataillon, G. Decoux, C. Rozale, D.J. Schoen, and J.L. David. 2001. MSTRAT: An algorithm for building germplasm core collections by maximizing allelic or phenotypic richness. J. Hered. 92:93–94. doi:10.1093/jhered/92.1.93

Hamblin, M.T., M.L. Warburton, and E.S. Buckler. 2007. Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. PLoS ONE 2(12):e1367. doi:10.1371/journal.pone.0001367

Jansen, J., and Th.J.L. van Hintum. 2007. Genetic distance sampling: A novel sampling method for obtaining core collections using genetic distances with an application to cultivated lettuce. Theor. Appl. Genet. 114:421–428. doi:10.1007/s00122-006-0433-9

Kaufman, L., and P.J. Rousseeuw. 1990. Finding groups in data. An introduction to cluster analysis. John Wiley & Sons, Hoboken, NJ.

Kim, K.W., H.K. Chung, G.T. Cho, K.H. Ma, D. Chandrabalan, J.G. Gwag, T.S. Kim, E.G. Cho, and Y.J. Park. 2007. PowerCore: A program applying the advanced M strategy with a heuristic search for establishing core sets. Bioinformatics 23:2155–2162. doi:10.1093/bioinformatics/btm313

Kirkpatrick, S., C.D. Gelatt, and M.P. Vecchi. 1983. Optimization by simulated annealing. Science 220:671–680. doi:10.1126/science.220.4598.671

Kover, P.X., W. Valdar, J. Trakalo, N. Scarcelli, I.M. Ehrenreich, M.D. Purugganan, C. Durrant, and R. Mott. 2009. A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. PLoS Genet. 5:E1000551. doi:10.1371/journal.pgen.1000551

Laval, G., M. San Cristobal, and C. Chevalet. 2002. Measuring genetic distances between breeds: Use of some distances in short term evolution models. Genet. Sel. Evol. 34:481–507. doi:10.1186/1297-9686-34-4-481

Lee, C., A. Abdool, C.H., Huang. 2009. PCA-based population structure inference with generic clustering algorithms. BMC Bioinformatics 10 (Suppl. 1):S73. doi:10.1186/1471-2105-10-S1-S73

Odong, T.L., J. van Heerwaarden, J. Jansen, Th.J.L. van Hintum, and F.A. van Eeuwijk. 2011. Determination of genetic structure of germplasm collections: Are traditional hierarchical clustering methods appropriate for molecular marker data? Theor. Appl. Genet. doi:10.1007/s00122-011-1576-x.

Paschou, P., E. Ziv, E.G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M.W. Mahoney, and P. Drineas. 2007. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet. 3:1672–1686. doi:10.1371/journal.pgen.0030160

Patterson, N., A.L. Price, and D. Reich. 2006. Population structure and eigenanalysis. PLoS Genet. 2:2074–2093. doi:10.1371/journal.pgen.0020190

Schoen, D.J., and A.H.D. Brown. 1993. Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. Proc. Natl. Acad. Sci. USA 90:10623–10627. doi:10.1073/pnas.90.22.10623

Thachuk, C., J. Crossa, J. Franco, S. Dreisigacker, M. Warburton, and G.F. Davenport. 2009. Core Hunter: An algorithm for sampling genetic resources based on multiple genetic measures. BMC Bioinformatics 10:243. doi:10.1186/1471-2105-10-243

Zhang, F., L. Zhang, and H.W. Deng. 2009. A PCA-based method for ancestral informative markers selection in structured populations. Sci. China C Life Sci. 52:972–976. doi:10.1007/s11427-009-0128-y

Zhang, H., D. Zhang, M. Wang, J. Sun, Y. Qi, J. Li, X. Wei, L. Han, Z. Qiu, S. Tang, and Z. Li. 2010. A core collection and mini core collections of *Oryza Sativa* L. in China. Theor. Appl. Genet. doi:10.1007/s00122-010-1421-7.