Theor Appl Genet (2013) 126:763–772
DOI 10.1007/s00122-012-2016-2

ORIGINAL PAPER

Maximizing genetic differentiation in core collections by PCA-based clustering of molecular marker data

Joost van Heerwaarden · T. L. Odong · F. A. van Eeuwijk

Received: 20 October 2011 / Accepted: 5 November 2012 / Published online: 21 November 2012

Abstract Developing genetically diverse core sets is key to the effective management and use of crop genetic resources. Core selection increasingly uses molecular marker-based dissimilarity and clustering methods, under the implicit assumption that markers and genes of interest are genetically correlated. In practice, low marker densities mean that genome-wide correlations are mainly caused by genetic differentiation, rather than by physical linkage. Although of central concern, genetic differentiation per se is not specifically targeted by most commonly employed dissimilarity and clustering methods. Principal component analysis (PCA) on genotypic data is known to effectively describe the inter-locus correlations caused by differentiation, but to date there has been no evaluation of its application to core selection. Here, we explore PCA-based clustering of marker data as a basis for core selection, with the aim of demonstrating its use in capturing genetic differentiation in the data. Using simulated datasets, we show that replacing full-rank genotypic data by the subset of genetically significant PCs leads to better description of differentiation and improves assignment of genotypes to their population of origin. We test the effectiveness of differentiation as a criterion for the formation of core sets by applying a simple new PCA-based core selection method to simulated and actual data and comparing its performance to one of the best existing selection algorithms. We find that although gains in genetic diversity are generally modest, PCA-based core selection is equally effective at maximizing diversity at non-marker loci, while providing better representation of genetically differentiated groups.

Communicated by G. Bryan.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-012-2016-2) contains supplementary material, which is available to authorized users.

J. van Heerwaarden (✉) · T. L. Odong · F. A. van Eeuwijk
Biometris, Wageningen UR, Wageningen, The Netherlands
e-mail: joost.vanheerwaarden@wur.nl

Introduction

Crop germplasm collections are important repositories of genetic diversity for plant breeding. The difficulties associated with the management and use of large numbers of accessions (Brown 1989) call for the formation of minimally redundant subsets, or core collections, that maximize the amount of represented genetic diversity (Frankel 1984; van Hintum et al. 2000). Increasingly, molecular markers are used to choose core sets based on aspects of genetic diversity such as pairwise dissimilarity, allelic richness, or heterozygostity (Bataillon et al. 1996; Franco et al. 2005; Jansen and van Hintum 2007; Thachuk et al. 2009; Odong et al. 2011b), either by maximizing these measures directly or by defining groups for subsequent stratified sampling of genotypes using clustering algorithms.

An implicit assumption shared by these methods is that markers are informative of genetic differences at important genes and traits (Schoen and Brown 1993). Although seemingly reasonable, this assumption is only met when markers and genes are physically linked or when there is differential genetic ancestry within the sample. Such differential ancestry may result from reproductive isolation between populations or from kinship structure (Sillanpää 2010). Both processes cause heterogeneity in allele frequencies within the gene pool (Astle and Balding 2009), which we will refer to here as genetic differentiation. Genetic differentiation translates directly into genome-

wide correlations in allelic state, which means that maximizing differentiation when sampling from the gene pool will simultaneously increase genetic differences at both markers and unlinked loci. Conversely, sampling within a homogeneous population means that unlinked loci are essentially independent and that markers will be uninformative of genetic differences at other genes.

Marker densities used for core selection are typically low and differentiation can be expected to be the cause of most marker–gene associations (Ohta 1982). The efficacy of core selection methods will therefore depend on the ability to capture genetic differentiation in marker data. Surprisingly, the extent to which commonly used dissimilarity and clustering methods adequately describe genetic differentiation in genotypic data has seldom been evaluated.

Simple indices of marker dissimilarity are likely to be relatively poor measures of genetic differentiation, as they largely reflect uninformative differences due to random sampling of alleles within populations. Recent work in population genetics suggests that reducing multi-locus data by Principal Component Analysis (PCA) provides a better description of differentiation, since genome-wide correlations between loci are effectively captured by the leading Principal components (PCs). Specifically, when genotypes are sampled from k differentiated populations, the first $k-1$ PCs reflect allele frequency differences between these populations while the remaining PCs relate to allelic sampling error only (Patterson et al. 2006). In addition, a direct relation between the level of differentiation and Euclidean distance along the relevant PCs (McVean 2009) further suggests that distance measures based on PCA may be appropriate for representing genetic differentiation.

PCA is now commonly applied to marker data (Patterson et al. 2006; Becquet et al. 2007; Tishkoff et al. 2009) and has been used for genetic clustering (Lee et al. 2009; van Heerwaarden et al. 2010, 2011), but so far there has been no systematic study of its potential as a basis for more effective core selection. Here, we analyze how genotypic core selection may be improved using PCA-based clustering to capture genetic differentiation in the data. We present our analyses for simple sequence repeat (SSR) data, currently the predominant data type in crop genetic resources. For simplicity, we assume that germplasm collections are samples from a limited number, k, of differentiated, genetically homogeneous populations of similar genetic diversity. We thereby focus on small core sets of k or fewer individuals, since under the aforementioned assumptions molecular markers contain no additional genetic information on repeated samples from the same populations.

Using simulated datasets, we measure to what extent the Euclidean distance along the first $k-1$ PCs improves the description of genetic differentiation over traditional dissimilarity indices that use full-rank genetic data. We then test if clustering based on this reduced set of PCs leads to better assignment of genotypes to their population of origin. To choose the number of PCs to retain, we apply a recent statistical criterion (Patterson et al. 2006; van Heerwaarden et al. 2010) that we adapt for use with SSR data. Finally, we demonstrate the effectiveness of PCA-based clustering for maximizing genetic dissimilarity within small core sets by comparing a simple PCA-based selection scheme to one of the best current core selection algorithms (Thachuk et al. 2009) using simulated and actual data (Odong et al. 2011a).

Methods

Example dataset

Our example dataset consists of 1,010 individuals of coconut (*Cocos nucifera* L.), sampled throughout the species geographic range and characterized for 28 SSR loci. The data were obtained from the central registry hosted by the Generation Challenge Programme—GCP (http://www.generationcp.org) and were chosen for its high data quality and the availability of a published analysis of genetic structure (Odong et al. 2011a).

Geographic origin of accessions, with abbreviations and number of individuals between parenthesis was as follows: Panama (CA: 104), Jamaica (CAR: 4), East Africa (EA: 118), West Africa (tall, WA: 29, dwarf varieties WAd: 3), Brazil (LA: 70), Pacific (tall, PCF: 343, dwarf varieties, PCFd: 14), South Asia (SA: 59), South East Asia (tall, SEA: 138, dwarf varieties, SEAd: 40), and Mexico (Pacific, NA1: 41 and Atlantic, NA2: 9).

Definition of genetic differentiation

The relation between genetic differentiation and inter-locus correlation in allelic state follows directly from the properties of the most common measure of differentiation, Wright's F_{st} (Wright 1951). F_{st} is a measure of both the allele frequency variance between reproductively separated populations and of the coancestry between gametes sampled from the same population. It therefore relates directly to the expected genetic differences between individuals from different populations and to the genetic correlation between unlinked loci. In the absence of gene flow, genetic drift will cause an increase in F_{st} according to the relation Time (generations) $= -\ln(1-F_{st})$ (Reynolds et al. 1983). In our study, we therefore defined pairwise genetic differentiation between populations i and j as $d_{F,i,j} = -\ln(1-F_{st,i,j})$ and use this measure in all our evaluations of

representation and maximization of genetic differentiation. $F_{st,i,j}$ was calculated according to Weir and Cockerham (1984) using a custom script in R (R, D.C.T 2009).

Population genetic simulations

Simulations were performed under the standard coalescent using an adapted version of the program msHOT (Hellenthal and Stephens 2007; Hudson 2002) (Code available on request). Multi-locus SSR genotypes were simulated under a stepwise mutation model, the simplest and most commonly used model of SSR evolution (Kimura and Ohta 1978). Thirty unlinked loci were simulated for a total of 1,100 haploid genotypes, divided into ten differently sized samples (20, 40…200 haploid genotypes) from ten discrete populations.

The population recombination rate ρ ($4N_e$c, where N_e is the effective population size and c the recombination fraction per generation) was set to 10,000 to guarantee independence between loci. The population mutation rate θ of 9 was set to achieve similar expected heterozygosity to that observed in our coconut dataset (He = 0.69). For applications requiring different levels of differentiation between population pairs, we used a random migration matrix with resulting pairwise F_{st} values between 0.06 and 0.15 (mean 0.11). For other applications, three fixed migration rates were used ($F_{st} = 0.06,\ 0.11,\ 0.20$) to achieve equal differentiation between populations. Levels of differentiation were chosen to be lower than those observed in our sample dataset. Diploid individuals were constructed by sampling random sets of two or four genotypes with replacement, yielding 550 individuals.

Capturing differentiation by principal component analysis of genotypic data

Genotypic dissimilarity is typically calculated on the full-rank matrix of allele counts. Here, we propose using a reduced set of PCs obtained by PCA of the original data matrix to accentuate between-population differentiation in the data (Lee et al. 2009).

We represent genotypic information by a set of individual by allele matrices $\mathbf{S_i}$ ($i = 1…L$), containing the allele counts for each allele of locus i. The number of columns of $\mathbf{S}_i$ is equal to the number of alleles per locus, with each column having a minimum integer value of 0 and a maximum value equal to the ploidy level (2 in the diploid case). Since the columns of $\mathbf{S}_i$ are not independent, we followed a normalization procedure that was previously proposed for linked SNP data (van Heerwaarden et al. 2010). Briefly, a normalized matrix $\mathbf{M}$ with independent columns of equal variance was created by concatenating PCs calculated separately for each $\mathbf{S_i}$ (using the function prcomp in R). After removing columns (PCs) that explain less than 0.5 % of total variance, each remaining column was standardized by dividing by its standard deviation.

Principal component analysis is subsequently performed on this normalized matrix $\mathbf{M}$. Since inter-locus correlations caused by differentiation between k populations should be fully represented by the first $k-1$ PCs, we considered only $k-1$ PCs in our evaluations of dissimilarity and clustering using simulated data, where k was known.

Correlation between dissimilarity measures and genetic differentiation

Genotypic core selection requires a measure of marker dissimilarity that adequately predicts dissimilarity at loci of agronomic interest. As discussed above, such inter-locus predictability depends on genetic differentiation, making it important to establish how well dissimilarity measures correlate with genetic differentiation. Differentiation is defined at the population level, with individuals within populations assumed to be unrelated. For k populations, the differentiation for all pairwise population comparisons is defined by a matrix of $k \times (k + 1)/2$ $d_{F,i,j}$ values, with a 0 diagonal. A matrix of genetic differentiation between n individual genotypes can therefore be represented by expanding the above matrix of size $k \times (k + 1)/2$ into a block matrix of $n \times (n-1)/2$ elements.

Using simulations with random migration (mean F_{st} 0.11, 10 replicates), we assessed the correlation of $d_{F,i,j}$ with the following pairwise dissimilarity measures: Euclidean distance based on the full matrix of allele counts (e.g. Reif et al. 2005), proportion of shared alleles (Bowcock et al. 1994); Nei 1972); Reynolds (Reynolds et al. 1983), 1-Jaccard (Jaccard 1908), 1-Dice (Dice 1945) and Rogers and Tanimoto (Rogers and Tanimoto 1960). We compared the above measures to what we will refer to here as PCA-reduced dissimilarity, the Euclidean distance calculated from the first $k-1$ PCs obtained from PCA on the normalized genotypic matrix $\mathbf{M}$.

We evaluated the correlation of dissimilarity measures with $d_{F,i,j}$ in two different ways. First, we measured the correlation between the matrix of $k \times (k-1)/2$, off-diagonal $d_{F,i,j}$ values, and a $k \times (k-1)/2$ matrix of mean dissimilarity values, obtained by averaging $n \times (n-1)/2$ pairwise values over their corresponding (among-population) blocks. This provides an estimate of the extent to which the mean value of each measure correlates to genetic differentiation between known populations. This is of theoretical relevance, since measures such as Nei and Reynolds dissimilarity were conceived as population-level estimators of genetic difference. Second, we estimated the correlation between the matrix of $n \times (n-1)/2$ individual pairwise dissimilarity values to the structured block matrix

of size $n \times (n-1)/2$ containing $k\ (k+1)/2$ blocks of $d_{F,i,j}$ values. This correlation estimates the capacity to capture differentiation without knowing the population delimitations and is of practical interest as it directly relates to subsequent clustering performance. Quantification of differences was done by ANOVA, followed by Fisher's protected least significant difference procedure at $p = 0.05$, after correction for replicate means. Untransformed values were used for convenience.

Ability of classification methods to capture genetic differentiation

Assuming that germplasm collections consist of discrete, internally homogeneous groups, genetic differentiation in the germplasm collection can be completely defined by classifying genotypes into distinct genetic clusters. The extent to which this adequately describes differentiation depends on the capacity to assign individuals to their correct genetic group.

Based on simulated datasets with three different levels of genetic differentiation ($F_{st} = 0.06$, 0.11, 0.20, 10 replicates), we compared different clustering methods for assignment success. We measured assignment success as the correlation between the two $n \times (n-1)/2$ pairwise indicator matrices of true and inferred shared group membership, coded as a binary state (1 for pairs within the same group, 0 for pairs from different groups). We compared the following methods: *k*-means, *k*-medoid (Kaufman and Rousseeuw 1990), UPGMA, and Ward. The *k*-means and medoids methods represent two commonly used non-hierarchical clustering techniques while UPGMA and Ward's algorithm are common hierarchical clustering techniques. We assumed the number of differentiated groups to be known, assigning genotypes to the $k = 10$ groups using the function cutree in R.

All methods were applied to a full-rank dissimilarity matrix, using the measure with the highest overall correlation to $d_{F,i,j}$, and to the PCA-reduced dissimilarity matrix. We also included model-based Gaussian hierarchical clustering (Fraley 1998) on the first $k-1$ PCs, as recently applied to human SNP data (Lee et al. 2009). This method assigns individuals to groups by fitting a mixture of multivariate normal distributions to the data and calculating the maximum likelihood of group membership by expectation maximization (EM) (Banfield and Raftery 1993). We used the implementation provided by the R package *mclust* (Fraley and Raftery 1999, 2002), employing the spherical, variable volume (VII) covariance model. This model was chosen because of the expectation of spherical clusters under an island model without admixture (McVean 2009). Statistical evaluation of differences was performed as described above.

Choosing the appropriate number of clusters

Although our evaluation of clustering performance assumes a priori knowledge of the number of differentiated groups, in practice the number of genetic groups needs to be inferred (Franco et al. 1997). It was shown recently (Patterson et al. 2006) that PCA on a normalized matrix of m independent (i.e., unlinked) SNP markers scored in n individuals, sampled from k differentiated populations and with $m > n$, yields $k-1$ significant eigenvalues when tested against a Tracy–Widom (TW) distribution (Tracy and Widom 1994; Johnstone 2001). Our normalized SSR genotype matrix $\mathbf{M}$ satisfies the requirement of independent columns of equal variance, and eigenvalues of covariance matrix of $\mathbf{M}$ approximately follow the TW distribution in the absence of structure (Van Heerwaarden et al. 2010). We follow the exact procedure of Patterson et al. 2006, but propose using the transposed matrix $\mathbf{M}^{\mathbf{T}}$ instead of $\mathbf{M}$ when $m < n$. This restores the condition of $n < m$ (Johnstone 2001) without affecting the actual eigenvalues and simulations show it provides a better fit to the TW distribution (Figure S1a, b). We compared this method to statistical stopping criteria based on Ward clustering of the pair-wise Euclidean distances. Ward clustering was recently shown to work better than UPGMA with most stopping criteria (Odong et al. 2011a). Using 100 simulated datasets of ten populations with three levels of differentiation ($F_{st} = 0.06$, 0.11, 0.20), we evaluated the following stopping criteria: Rousseeuw's Silhouette internal cluster quality index, Point biserial index, Hubert and Levin C-index (Milligan and Cooper 1985; Odong et al. 2011a).

Using PCA-based clustering to maximize differentiation during core selection

Once differentiated populations within the germplasm collection have been correctly defined, we may in principle ignore marker dissimilarities within these populations and aim directly at maximizing genetic differentiation between represented genetic groups. To demonstrate this approach, we explore the case of a small core set, where the size of the subset is equal or smaller than the actual number of differentiated groups. For this case, we propose the following simple procedure for genotype selection using PCA-based clustering:

(1) Do PCA on the normalized genotype matrix $\mathbf{M}$ as described above. (2) Determine the number of significant eigenvalues, say $k-1$, of $\mathbf{M}$ (or $\mathbf{M}^{\mathrm{T}}$ when $m < n$) by comparing the value of normalized eigenvalues to the TW distribution. (3) Using *mclust,* assign individuals to k groups by fitting a mixture of k multivariate normal distributions to the matrix of $k-1$ significant PCs obtained

from PCA on **M**. This yields the k-dimensional multivariate means (center points) of each group. (4) For a desired genotype subset of size s (with $s < k$), generate all possible samples of genetic groups and calculate the mean Euclidean distances in $k-1$ dimensional PC space between the center points of the sampled groups. This will be computationally feasible up to $k \sim 35$. Choose the groups that maximize the Euclidean distance between their center points, and select the s individuals with the least Euclidean distance to center point of their corresponding group. This procedure is aimed at maximizing genetic differentiation between the populations from which the selected genotypes are sampled and will work as long as k is relatively small.

Validation of PCA-based core selection by comparison with Core Hunter

We evaluated the performance of this PCA-based core selection procedure by comparing against Core Hunter (Thachuk et al. 2009), a recent stochastic local search (SLS) algorithm that selects genotypes by maximizing one or several genetic distance and diversity measures by replica exchange Monte Carlo. It was shown by its authors to outperform several existing core selection methods, particularly for maximizing single diversity measures at marker loci. We simulated 10 populations with fixed migration rates to ensure identical differentiation for each population (50 replicates). At three levels of differentiation ($F_{st} = 0.06, 0.11, 0.2$), we tested the performance of both methods in maximizing the mean Euclidean distance at 30 marker loci and 30 target loci (i.e., functionally important non-marker loci) in a subset of 10 selected genotypes. Core Hunter was thereby set to maximize the Modified Rogers distance, which is proportional to the Euclidean distance (Reif et al. 2005). We also evaluated the number of represented populations, as well as maximization of average pairwise F_{st} between sampled populations, for the case where the number of populations exceeded number of selected individuals. The latter was evaluated in core sets of five individuals selected from 10 differentiated populations (random migration, pairwise $F_{st} = 0.06$–0.15, 20 replicates).

For the coconut data, we compared a small subset of 5 individuals selected by Core Hunter to a subset chosen by our core selection procedure. We removed highly inbred individuals by setting a cut-off at the 0.05 quantile of observed heterozygosity. Performance at target loci was evaluated by selecting based on random subsets of 20 loci and measuring the Euclidean distance at the 8 remaining loci. Significance between the two methods and random selection was tested with ANOVA as described above, using 100 sub-samples as replicates. We also compared the realized population differentiation, d_F, within each subset, where d_F was calculated between genetic groups identified by model-based Gaussian clustering on the significant PCs.

Results

Correlation between dissimilarity measures and genetic differentiation

Reynold's dissimilarity has the highest correlation with genetic differentiation (Fig. 1, Table S1) when averaged over among-population comparisons, reflecting the fact that its theoretical expectation is F_{st} (Reynolds et al. 1983). It performs poorly as an individual-level dissimilarity, however, which is probably due to its large variance (Reynolds et al. 1983). Jaccard dissimilarity, proportion of shared alleles, and Euclidean distance all perform significantly better in this respect. In addition, the among-population means of the Euclidean distance also has a high correlation with d_F, suggesting it as the most appropriate full-rank dissimilarity for describing genetic differentiation. Although the among-population means of PCA-reduced dissimilarity do not have a particularly high correlation with d_F, it clearly outperforms all other dissimilarities as an individual measure, confirming its potential for capturing genetic differentiation in the absence of prior population definitions.

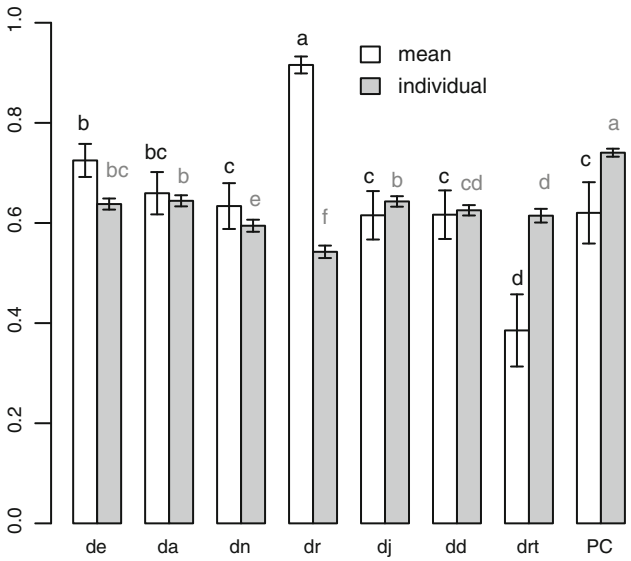

Fig. 1 Correlation between different distance measures and pairwise population differentiation d_F, in ten simulated datasets ($F_{st} = 0.11$). Results are presented for individual distances and distances averaged over population comparisons. *de* Euclidean, *da* proportion of shared alleles, *dn* Nei (1972), *dr* Reynolds, *dj* Jaccard, *dd* Dice, *drt* Rogers & Tanimoto, *PC* Euclidean distance using $k-1$ PCs. *Whiskers* represent standard errors. *Letters* above each bar indicate significantly different means within each of the two distance types (Fisher's LSD)

Ability of classification methods to capture genetic differentiation

Clustering performance, as measured by the correlation between true and inferred population assignment matrices, confirms the benefit of reducing genetic data by PCA (Fig. 2, Table S2). At every level of differentiation, all clustering methods show improved assignment success when using PCA-reduced dissimilarity instead of the full-rank Euclidean distance matrix (ANOVA, $p < 0.0001$). Although at the highest levels of differentiation performance deviates little between methods, at lower levels differences become evident. UPGMA clustering is clearly the weakest method, while Ward clustering using PCA-reduced dissimilarity and particularly model-based clustering on the first $k-1$ PCs outperform the other methods in terms of assignment success.

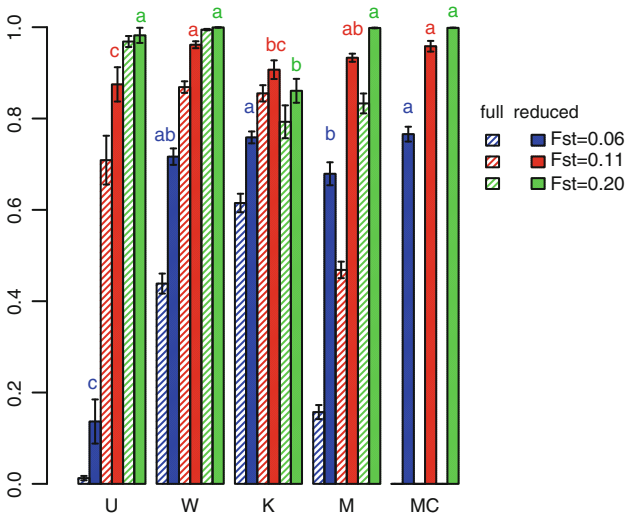

Fig. 2 Assignment success (correlation between true and inferred binary assignment matrices) in ten simulated datasets at three levels of differentiation, using the full-rank dissimilarity matrices (full) and PCA-reduced similarity (reduced). *U* UPGMA, *W* Ward K: *k*-means, *M* Medoid clustering, *MC* model-based clustering on 9 PCs, *Whiskers* represent standard errors. *Letters* above each bar indicate significantly different means within each level of differentiation (Fisher's LSD)

Choosing the appropriate number of clusters

Although at the highest level of differentiation all stopping criteria show similar (Fig. 3) performance, TW-based PC significance is the only criterion that provides reasonable estimates of the true number of genetic groups across all levels of differentiation. The Point Biserial index and particularly the C-index considerably overestimate the number of groups at lower levels of differentiation, while Rousseeuw's Silhouette index quite severely underestimates the number of groups at $F_{st} = 0.06$. In spite of its stable performance, TW-based PC significance also underestimates the number of groups by 1 and 2 at the two lower levels of differentiation, probably reflecting a failure to detect the smallest populations.

Validation of PCA-based core selection by comparison with Core Hunter

Core Hunter clearly outperforms PCA-based core selection in terms of maximization of dissimilarity at marker loci (Fig. 4). For each level of differentiation, the average full-rank Euclidean distance over the 50 replicates was about 16–21 % higher than that achieved by PCA-based and random selection. At target (non-marker) loci, however, the average distance is only slightly higher than that of random and equal to that achieved by PCA-based selection. Both selection methods achieve distance gains that, although significant, are only 2 % above random expectations.

In terms of representation of genetic differentiation, defined as the number of sampled populations, selection by PCA-based clustering outperforms Core Hunter at all three levels of differentiation. At $F_{st} = 0.06$, it samples 9 or more groups in 68 percent of cases (average 8.7) against no cases for Core Hunter (average 6.8). For $F_{st} = 0.11$ and $F_{st} = 0.20$, PCA-based selection yields 9 or more groups in

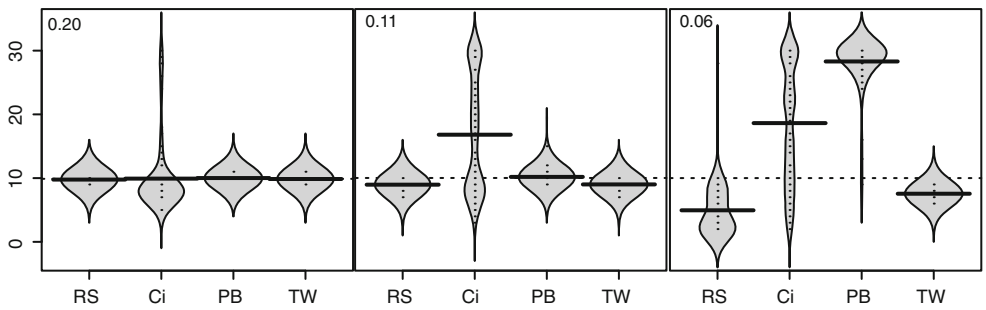

Fig. 3 Violin plots showing the density distribution of the number of groups (*vertical axis*) inferred for 100 simulated datasets, using different stopping criteria. Results are presented for three levels of differentiation (F_{st}: 0.20, 0.11, 0.06). *RS* Rousseeuw's Silhouette internal cluster quality index, *Ci* Hubert and levin C-index, *PB* Point biserial index, *TW* Tracy-Widom significance

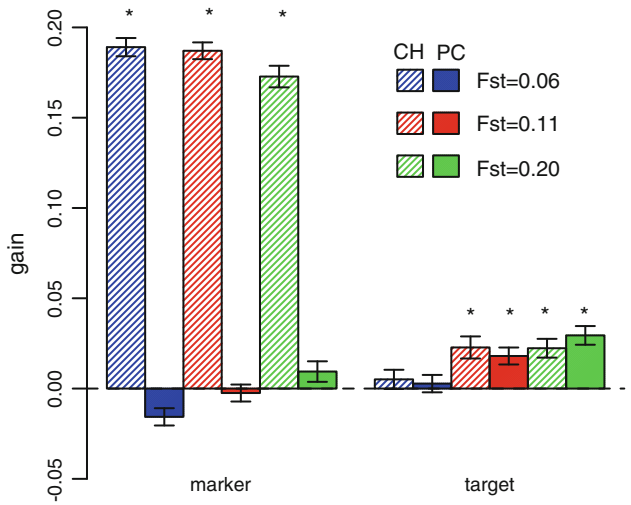

Fig. 4 Proportional gain in mean Euclidean distances of selected over random core sets (10 individuals), sampled from ten simulated populations at three levels of differentiation. *PC* selection by PCA-based clustering, *CH* selection based on Core Hunter, *marker* marker loci, *target* target (non-marker) loci. *Stars* indicate values significantly higher than 0

96 and 88 % of cases (average 9.4 and 9.3) against 6 and 38 % (average 7.4 and 8.2) for Core Hunter. Both methods perform better than random. Similarly, PCA-based selection of a subset of five populations achieves a significantly higher differentiation of $d_F = 0.211$ (maximum achievable value of 0.222) against 0.201 for Core Hunter and 0.179 for random selections.

PCA-based clustering and core selection in coconut

PCA on the coconut dataset reveals the presence of significant population structure. The total number of significant PCs is 14, explaining 23 % of variance. Model-based clustering into 15 groups produces geographically sensible clusters (Fig. 5), with an average pairwise d_F of 0.25. Some clusters are closely related and form larger geographic groups. The Pacific accessions (PCF) are represented by as many as 6 clusters, some of which show differentiation as high as 0.14. The Panamanian (CA) accessions form a rather distinct set of two clusters that group relatively close to South East Asian (SEA) and Pacific (PCF) accessions. Another set of three closely related clusters clearly separate the South Asian (SA), Brazilian (LA), and West African (WA) accessions. Consistent with earlier results (Odong et al. 2011a), Mexican Pacific (NA1) and Atlantic accessions fall in different clusters with Pacific (PCF) and South Asian (SA) accessions, respectively. The South East Asian Dwarf accessions (SEAd) also form a separate cluster, together with a number of Pacific Dwarf (PCFd) accessions.

Comparison between our PCA-based core selection method and Core Hunter yields similar results to the simulated data. Core Hunter greatly outperforms PCA-based selection at marker loci, with a mean Euclidean distance that is 24 % higher (0.82 against 0.66) and 35 % better than random (0.61). At target (non-marker) loci, a much lower although significant 5 % difference remains (0.70 vs. 0.67) with the two methods performing 16 and 12 % better than random (0.61). Inspection of the core selections produced by both methods reveals that contrary to PCA-based core selection, Core Hunter tends to select the same genotypes repeatedly. Six individuals in particular are sampled more than 25 out of 100 times by Core Hunter. Low levels of observed heterozygosity in these individuals (0.19 compared to an average of 0.43) suggest that these may be relatively inbred genotypes acting as outliers with respect to overall patterns of differentiation. Excluding these genotypes confirms this. Although Core Hunter still achieves substantial gains at marker loci (0.80 against 0.66 in the PCA-based selection), the difference at target loci is no longer significant (0.67 against 0.68).

Discussion

Core collections need to capture the maximum amount of genetic diversity in a small subset of accessions (Frankel 1984). Presently, molecular markers are the most popular means of describing this diversity, either by themselves or in conjunction with phenotypic traits (Schoen and Brown 1993; Franco et al. 2009). In contrast to phenotypic traits, unlinked molecular markers are not of interest in themselves, but serve to detect genetic differentiation associated with patterns of reproductive isolation within the gene pool. Since differentiation determines the level of genetic correlation between loci, it can be argued to provide the best predictor of genome-wide patterns of dissimilarity. In spite of its importance, most methods of marker-based core selection do not take explicit account of genetic differentiation but instead focus on maximizing dissimilarity at marker loci (e.g. Franco et al. 2006; Thachuk et al. 2009).

We have shown here that most dissimilarity measures are relatively poor descriptors of genetic differentiation, even when their expected values correlate strongly with d_F. This is not entirely surprising, as their within-population variance is high due to the small sample sizes implied by the comparison of individual genotypes (Nei and Roychoudhury 1974; Goldstein et al. 1995).

We confirm that representing the genotypic data by the subset of PCs associated with genetic differentiation provides a better basis for describing genetic structure in the data. The Euclidean distance calculated on these PCs has a stronger correlation with genetic differentiation than full-rank distance measures, while improving population assignment success. In agreement with a recent study using

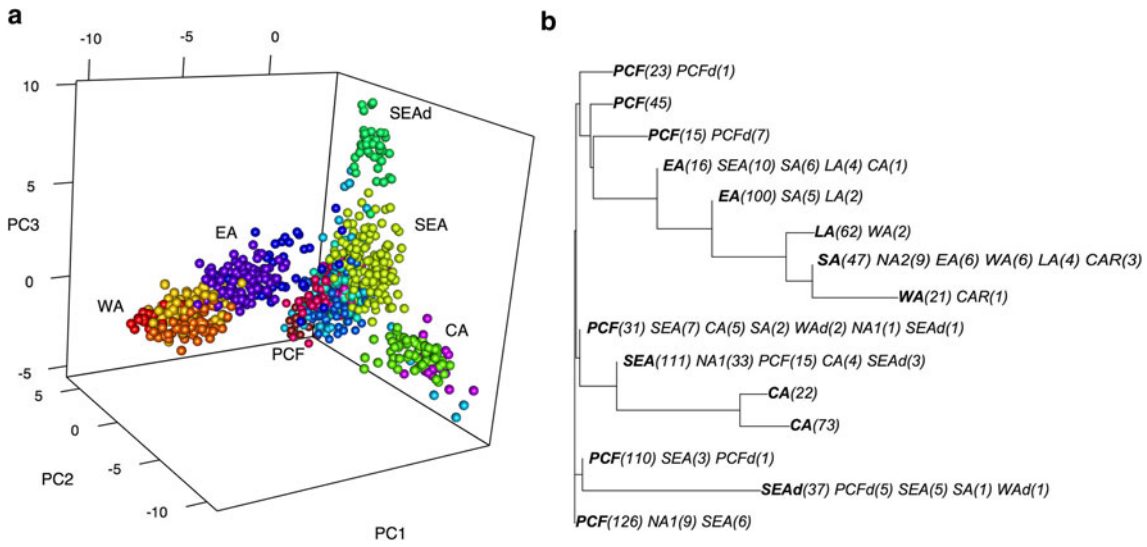

Fig. 5 PCA plot with accessions colored by genetic group (**a**) and Neighbor Joining tree based on pairwise d_F between the 15 PCA groups (**b**). *CA* Panama, *CAR* Jamaica, *EA* East Africa, *WA* West Africa (tall), *Wad* West Africa (dwarf), *LA* Brazil, *PCF* Pacific (tall), *PCFd* Pacific (dwarf), *SA* South Asia, *SEA* South East Asia (tall), *SEAd* South East Asia (dwarf), *NA1*: Mexico (Pacific), *NA2* Mexico (Atlantic)

SNP data (Lee et al. 2009), we find that the combination of PCA with model-based clustering using *mclust* performs particularly well. These results prove the potential of PCA-based clustering for core set selection with the aim of representing genetic differentiation.

The practical application of PCA-based clustering requires an effective method for defining the number of components to be retained. Although many statistical criteria exist (Milligan and Cooper 1985), none relates directly to genetic differentiation. We demonstrate that testing the significance of eigenvalues against the TW distribution (Tracy and Widom 1994), as pioneered by Patterson et al. (2006), works well for properly normalized SSR data. Significant PCs reflect the covariance between markers that is caused by differential coancestry in the data (Patterson et al. 2006) thus providing a statistic that relates directly to genetic differentiation and the expected correlation between unlinked loci. The value of this procedure is evident not only from its favorable performance on simulated data but also by its ability to reveal a larger number of geographically sensible groups than found in earlier studies (Odong et al. 2011a).

The capacity of PCA-based clustering to capture genetic differentiation suggests that it provides a powerful basis for the selection of genotypic core sets. This is confirmed by the comparison of our implementation of PCA-based core selection with the Core Hunter algorithm (Thachuk et al. 2009). Core Hunter was reported to outperform two of the most widely used core selection methods, *Mstrat* (Gouesnard et al. 2001) and Franco et al.'s D method (Franco et al. 2005) in creating sets of maximum diversity. Although in our simulations Core Hunter indeed achieves the highest average distance at marker loci, its performance at non-marker loci does not surpass that achieved by our method. In addition, by ignoring genetic differentiation, Core Hunter tends to under-sample differentiated populations. We show that selection using PCA-based clustering achieves similar gains at target loci while better presenting the genetic differentiation in the data.

These results show that although maximization algorithms such as Core Hunter may be effective in increasing dissimilarity, the actual gains will not exceed what can be achieved by adequately sampling differentiated populations. In fact, the consideration that within homogenous populations, unlinked markers are genetically uninformative, means that in many cases the problem of high dimensionality (i.e., the large number of pairwise relationships to consider) in core selection can be reduced by PCA-based clustering. For the small core sets considered here, rather than considering all possible combinations of sampled individuals, accessions may be chosen by maximizing differentiation between a limited number of internally homogeneous groups. In addition, since samples from the same group are genetically equivalent under our assumptions, larger core sets can simply be obtained by uniform stratified sampling, although sampling weighted by within-group diversity may be more desirable in practice (Franco et al. 2005). Our procedure obviously works best when a limited number of well sampled groups are present in the data. Our observation of the effect of outlier individuals in the coconut data, for example, suggests that actual data may sometimes deviate from these simple assumptions. For very complex coancestry patterns or when the data contain many groups, it may therefore be

preferable to use an algorithm like that implemented in Core Hunter to maximize PCA-based distances directly, rather than relying on the identification of discrete clusters.

Finally, it is interesting to note that in spite of the good performance of our core selection method, gains in both dissimilarity at target loci and genetic differentiation are modest. Although seemingly discouraging, it does not reflect the limitations of our method but rather the realities of marker-based core selection. The number of differentiated groups and levels of differentiation typically present in the data mean that true gains will be modest, regardless of the method used. It is important to realize, however, that genetic PCs, by reflecting barriers to gene flow, often correlate with geography and environment (Manel et al. 2007; Eckert et al. 2010). As such, core selection using PCA-based clustering may allow better sampling of selective environments with potentially strong effects on genes and traits of agronomical interest.

Acknowledgments The authors wish to thank Carmen de Vicente, former leader of subprogram 5 of the Generation Challenge Program (GCP), for providing financial support (GCP 4008.23) and guidance. We thank Diego Ortega Del Vecchyo for contributing software and three anonymous reviewers for comments on earlier versions of the manuscript.

References

Astle W, Balding DJ (2009) Population structure and cryptic relatedness in genetic association studies. Stat Sci 24:451–471

Banfield JD, Raftery AE (1993) Model-based gaussian and non-gaussian clustering. Biometrics 49:803–821

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

Becquet C, Patterson N, Stone AC, Przeworski M, Reich D (2007) Genetic structure of chimpanzee populations. PLoS Genet 3:617–626

Bowcock AM, Ruizlinares A, Tomfohrde J et al (1994) High-resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455–457

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

Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26:297–302

Eckert AJ, van Heerwaarden J, Wegrzyn JL et al (2010) Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). Genetics 185:969–982

Fraley C (1998) Algorithms for model-based Gaussian hierarchical clustering. SIAM J Sci Comput 20:270–281

Fraley C, Raftery AE (1999) MCLUST: software for model-based cluster analysis. J Classif 16:297–306

Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 97:611–631

Franco J, Crossa J, Diaz J et al (1997) A sequential clustering strategy for classifying gene bank accessions. Crop Sci 37:1656–1662

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

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

Franco J, Crossa J, Desphande S (2009) Hierarchical multiple-factor analysis for classifying genotypes based on phenotypic and genetic data. Crop Sci 50:105

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

Goldstein DB, Linares AR, Cavallisforza LL, Feldman MW (1995) An evaluation of genetic distances for use with microsatellite loci. Genetics 139:463–471

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

Hellenthal G, Stephens M (2007) msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots. Bioinformatics 23:520–521

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338

Jaccard P (1908) Nouvelles recherches sur la distribution florale. Bull Soc Vaud Sci Nat 44:223–269

Jansen J, van Hintum T (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

Johnstone IM (2001) On the distribution of the largest eigenvalue in principal components analysis. Ann Stat 29:295–327

Kaufman L, Rousseeuw PJ (1990) Finding groups in data. An introduction to cluster analysis. Wiley, New York

Kimura M, Ohta T (1978) Stepwise mutation model and distribution of allelic frequencies in a finite population. Proc Natl Acad Sci USA 75:2868

Lee C, Abdool A, Huang CH (2009) PCA-based population structure inference with generic clustering algorithms. BMC Bioinform 10(Suppl 1):S73

Manel S, Berthoud F, Bellemain E et al (2007) A new individual-based spatial approach for identifying genetic discontinuities in natural populations. Mol Ecol 16:2031–2043

McVean G (2009) A genealogical interpretation of principal components analysis. PLoS Genet 5:e1000686

Milligan GW, Cooper M (1985) An examination of procedures for determining the number of clusters in a data set. Psychometrika 50:159–179

Nei M (1972) Genetic distance between populations. Am Nat 106:283

Nei M, Roychoudhury AK (1974) Sampling variances of heterozygosity and genetic distance. Genetics 76:379

Odong TL, van Heerwaarden J, Jansen J, van Hintum TJ, van Eeuwijk FA (2011a) Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data? Theor Appl Genet 123:195–205

Odong TL, van H J, Jansen J, van H TJL, van E FA (2011b) Statistical techniques for defining reference sets of accessions and microsatellite markers. Crop Science 51:2401

Ohta T (1982) Linkage disequilibrium with the island model. Genetics 101:139

Patterson N, Price AL, Reich D (2006) Population structure and eigen analysis. PLoS Genet 2:e190

R, DCT (2009) R: a language and environment for statistical computing

Reif JC, Melchinger AE, Frisch M (2005) Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. Crop Sci 45:1–7

Reynolds J, Weir BS, Cockerham CC (1983) Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105:767

Rogers DJ, Tanimoto TT (1960) A computer programming for classical plants. Science 132:1115–1118

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

Sillanpää MJ (2010) Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. Heredity 106:511–519

Thachuk C, Crossa J, Franco J et al (2009) Core Hunter: an algorithm for sampling genetic resources based on multiple genetic measures. BMC Bioinform 10:243

Tishkoff SA, Reed FA, Friedlaender FR et al (2009) The genetic structure and history of Africans and African Americans. Science 324:1035–1044

Tracy CA, Widom H (1994) Level-spacing distributions and the airy kernel. Commun Math Phys 159:151–174

Van Heerwaarden J, Ross-Ibarra J, Doebley J et al (2010) Fine scale genetic structure in the wild ancestor of maize (Zea mays ssp. *parviglumis*). Mol Ecol 19:1162–1173

van Heerwaarden J, Doebley J, Briggs WH et al (2011) Genetic signals of origin, spread, and introgression in a large sample of maize landraces. Proc Natl Acad Sci USA 108:1088–1092

Van Hintum TJL, Brown AHD, Spillane C, Hodgkin T (2000) Core collections of plant genetic resources. Bioversity International

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370

Wright S (1951) The genetical structure of populations. Ann Eugen 15:323–354