Neutral Genetic Markers and Conservation Genetics: Simulated Germplasm Collections

Thomas M. Bataillon,*,† Jacques L. David* and Daniel J. Schoen†

**Laboratoire INRA-ENSAM d'Amélioration des Plantes, 34100 Montpellier Cedex 01, France and †Department of Biology, McGill University, Montréal, Québec, H3A 1B1, Canada*

Manuscript received January 30, 1996
Accepted for publication June 11, 1996

ABSTRACT

This study examines the use of neutral genetic markers to guide sampling from a large germplasm collection with the objective of establishing from it a smaller, but genetically representative sample. We simulated evolutionary change and germplasm sampling in a subdivided population of a diploid hermaphrodite annual plant to create an initially large collection. Several strategies of sampling from this collection were then compared. Our results show that a strategy based on information obtained from marker genes led to retention of the maximum number of neutral and nonneutral alleles in the smaller sample. This occurred when demes were composed of self-fertilizing individuals or when no migration occurred among demes, but not when demes of an outcrossing population were connected by high levels of migration.

INCREASING habitat destruction in this century has raised concerns about genetic depletion in natural populations (WILSON 1992). Loss of genetic diversity is also a problem in the case of agriculturally important species, where ancient cultivars (or landraces) and wild relatives of domesticated species are being lost as modern varieties become adopted by farmers. This has led to calls for genetic conservation of crop germplasm (FRANKEL and BENNET 1970).

It is seldom if ever possible to assess in a comprehensive manner the amount and structure of genetic variation in a population or collection of interest to genetic conservation, making it difficult to proceed in a rational way toward construction of representative samples for conservation. A recent approach to this problem proposes the use of easily scorable genetic markers such as allozymes and DNA level polymorphisms to determine how single locus variation is structured among and within populations (BROWN and CLEGG 1983). SCHOEN and BROWN (1993) showed that if marker gene data are available for many populations of a species, significant gains in the number of neutral alleles retained in a sample can be achieved through a combination of stratified sampling and weighting of the contributions of the different populations to the germplasm collection according to information provided by the marker loci. But SCHOEN and BROWN (1993) tested their methods using allozyme loci, not only as a means to guide the sampling, but also as way to score the allelic richness of different samples. Because many allozyme polymorphisms are likely to be selectively neutral, it is unclear from their results whether a marker-based approach to genetic conservation would also lead to significant gains in the capture of adaptive genetic variation (HOLSINGER 1991; MILLIGAN *et al.* 1994). Here, we address this problem using computer simulation to assess the efficacy of different sampling strategies to capture both neutral and nonneutral allelic variation under a number of different ecological conditions.

A large number of materials from major crop species and their wild relatives are currently stored in international networks of seed banks, or in "*in situ*" conservation sites and "on farm" programs of conservation. Because the large size of some of these collections, together with limited funding, combine to restrict the characterization of the material available and hinder their use for breeding purposes (FRANKEL 1989; BROWN 1995), an increasingly popular proposal for germplasm management is to construct smaller "core collections" from these larger collections. Ideally, core collections should be chosen to represent the bulk of the genetic diversity contained in the larger collection. Construction of a such a collection usually starts by stratifying the larger collection into a series of groups, to acknowledge that accessions originate from different ecogeographic regions with possibly divergent evolutionary histories. The core collection is then established by stratified random sampling from the different groups (BROWN 1989a,b). While the work presented below is directed to the specific problem of how information gained from neutral markers can lead to increases in the amount of nonneutral (and neutral) variation in core collections, the results are relevant to other problems in conservation genetics where priorities for conservation among a set of different populations must be set.

Corresponding author: Daniel J. Schoen, Department of Biology, McGill University, 1205 Avenue Docteur Penfield, Montréal, PQ, H3A 1B1, Canada. E-mail: dan_schoen@maclan.mcgill.ca

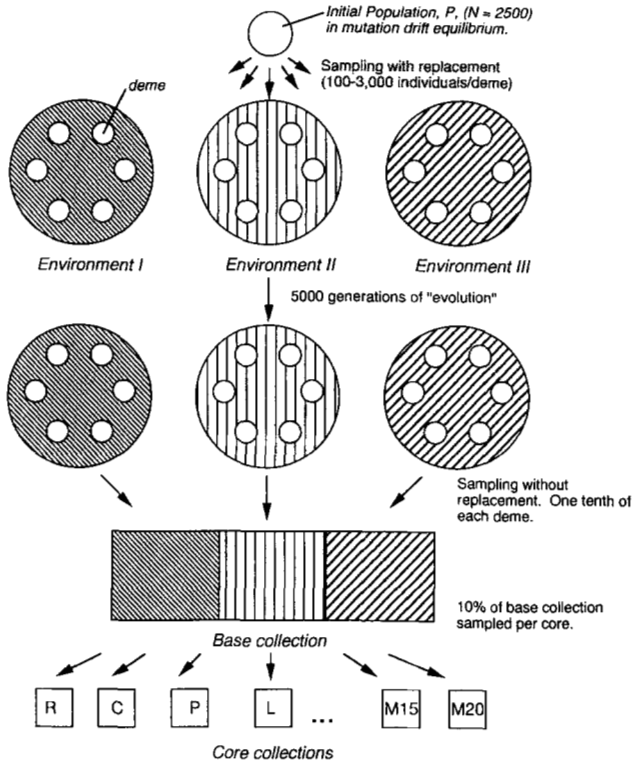

FIGURE 1.—Overview of the simulation and sampling algorithms. An initial population (Population P) consisting of 2500 diploid individuals, in mutation-drift equilibrium (at each of 99 loci) was sampled with replacement to create 18 demes, distributed across three selective environments. When simulating selfing populations, individuals in Population P were self-fertilized for 100 generations to form Population P_s, before populating the 18 demes. An equal number of demes per environment is depicted here (uneven distributions were also used in some simulations—see text). After 5000 generations of mating, drift, selection, and in some cases, migration (see text and Table 1), the 18 demes were sampled to form the base collection. Various sampling strategies were then used to construct the different core collections by sampling 10% of the base collection.

MATERIALS AND METHODS

Overview: The simulation and sampling algorithm is depicted in Figure 1. We simulated phenotypic evolution in a subdivided plant population distributed across three environments (denoted I, II, and III). This subdivided population contained all the available genetic variation. Once a quasi-equilibrium was reached between drift, mutation, migration and selection, we simulated the sampling of the population, leading to the formation of a large collection, hereafter referred to as the "base collection". In crop germplasm conservation, the base collection represents the stage of sampling preceding the creation of the core collection. Before constructing the core collection, the base collection was structured into three groups, reflecting the different environments of origin of the accessions. Finally, we simulated the construction of core collections from this base collection using different sampling strategies. These steps are described in more detail below.

Simulation of the plant population: A population of 2500 individuals (Population P), initially monomorphic at all 99 simulated loci, was used to start the simulation. Mutation, genetic drift, and gamete formation were simulated in this population for 15,000 generations. For each locus, it was assumed that any of 1000 different alleles could arise with equal probability as a result of recurrent mutation. A Poisson distribution of crossing over with no interference was assumed when simulating gamete formation (HOSPITAL and CHEVALET 1996). Following the 15,000 generations of mutation and drift in Population P, the allele frequency distributions at these loci were found to conform to expectation under selective neutrality, as verified by the Ewens-Watterson test (EWENS 1979; ENDLER 1986). In the case of simulations involving self-fertilization, an inbred population (Population P_s) was established by simulating 100 generations of complete self-fertilization in Population P (after the initial 15,000 generations of mutation and drift).

Eighteen separate demes were then established by sampling individuals with replacement from Population P (or Population P_s, in the case of selfing). Deme size ranged from 100 to 3000 diploid individuals, with a mean of 1000 individuals. The decision to reduce the effect of drift led to a minimum local population size of 100 individuals. Next, these demes were subjected to mutation, drift, migration and selection, for 5000 generations, using the simulation procedure of DAVID *et al.* (1993), modified to include migration among demes. An island model of gene flow allowing movement of diploid genotypes (*i.e.*, seed migration but not pollen migration) was assumed. The structure of the subdivided population remained constant throughout the simulation, and no extinction of demes was allowed. The subdivided population so constructed was intended to simulate a naturally occurring species, such as a wild crop relative.

Two groups of loci were simulated: selectively neutral loci (hereafter referred to as "marker loci") and loci under selection, in which fitness contributions were determined by the combination of genotype and environment (Figures 2 and 3). Each environment was characterized by a different selection regime, *i.e.*, the alleles at each locus were subject to different selection pressures in each of the three environments (Figure 2). Within environments, each selected locus contributed equally and additively to fitness.

A number of other features were kept constant throughout the simulation, including linkage relationships among the loci (Figure 3), the mutation rate per locus ($\mu = 0.0001$), and patterns of selection at the different loci in the three environments (Figure 2). The choice of a relatively high mutation rate was dictated in part by the need to produce high allelic diversity ($\theta = 4N_d\mu$) in the populations. This allowed detection of differences among the sampling strategies. Such mutation rates are characteristic of microsatellite DNA markers (HENDERSON and PETES 1992).

Eight different cases were studied by combining three features of the simulation in different ways. These features included: the mating system of the plants—selfing rate of 0 *vs.* 100%; the extent of gene flow—2% individuals from each deme allowed to migrate *vs.* complete isolation between demes; and distribution of the 18 demes across the three different environments—equal *vs.* variable numbers of demes per environment (Table 1). Each combination was replicated twice. These different combinations of conditions were intended to capture some of the essential population biology of species differing in mating systems and dispersal abilities (*e.g.*, self- and cross-fertilization, animal *vs.* gravity dispersal of seed), as well as the imbalance that often exists in base collections with regard to the number of samples originating from particular regions.

Simulation of the base collection: After 5000 generations of mutation, migration, drift and selection in the subdivided population of 18 demes, as described above, one-tenth of the individuals present in each deme were sampled to form the

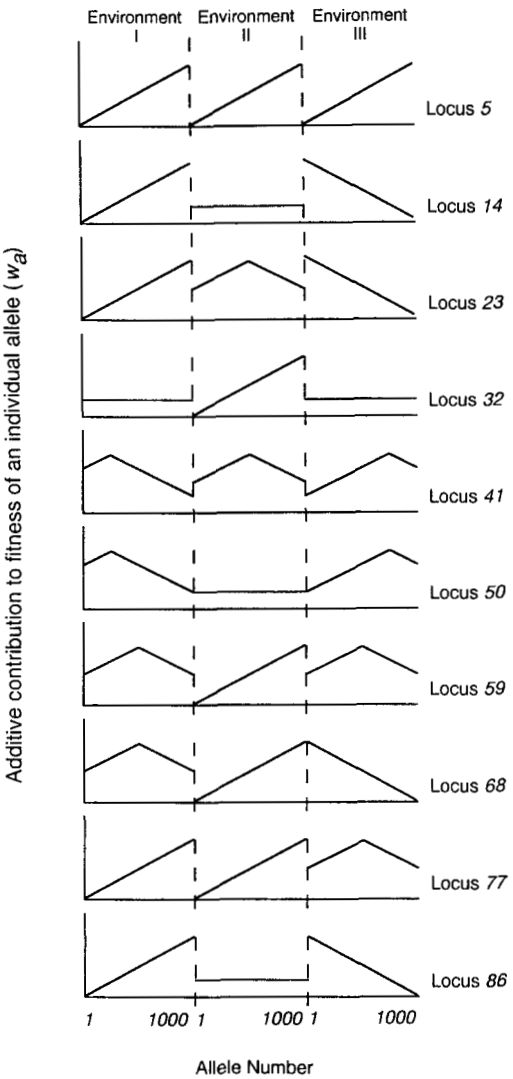

FIGURE 2.—Contributions to individual fitness of each of 1000 alleles at 10 loci under different modes of selection. Within each environment, individual fitnesses were determined by summing contributions (w_as) across loci. Dominance and epistasis were assumed to be absent.

base collection (Figure 1). These individuals were kept together in sets of 10 each. Each set of 10 individuals is referred to below as an "accession". This procedure parallels the practice in germplasm management of collecting and storing several seeds or plants per population as a means to represent the genetic composition of the population. The number of accessions, n_i, sampled from each deme i was $n_i = N_i/100$, where N_i denotes the number of individuals in the deme i. There were $n = \Sigma n_i$ accessions in the entire base collection (summation over all 18 demes). The base collection was organized into three groups by keeping accessions from the each of the three environments together. These groups of accessions are hereafter referred to as "diversity groups".

Two replicate base collections were made for each combination of mating system, migration rate, and deme distribution, leading to eight combinations of simulation conditions × two replicates per combination × two base collections per replicate for a total of 32 base collections simulated in total.

Comparison between core collection sampling strategies: Base collections were sampled to form core collections consisting of $r = 0.1 \times n$ accessions. Several core collection sampling strategies were used and are referred to as the R, C, P, L and M strategies (Figure 1, Table 2). The R strategy assumes no prior knowledge about the base collection (except for its total size) and involves only random sampling of accessions,

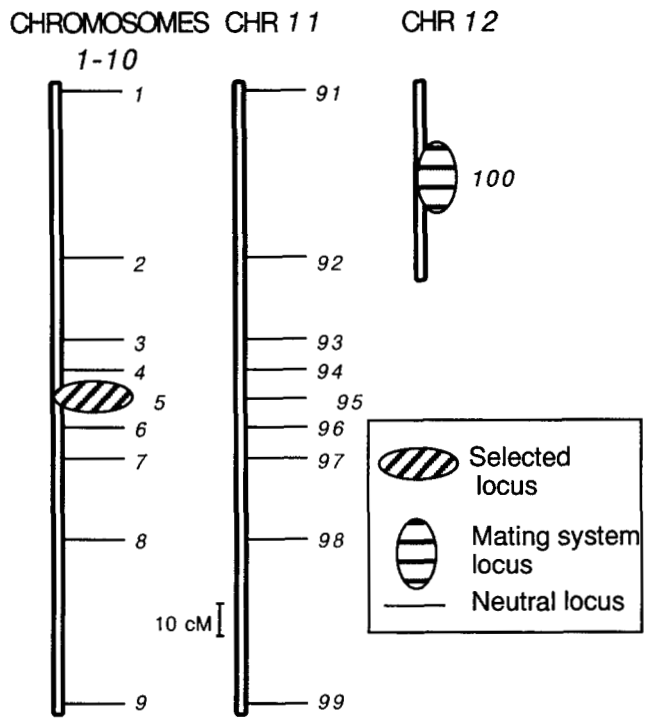

FIGURE 3.—Positions of neutral and selected loci among chromosomes in the simulation. Ten chromosomes with equivalent structure, plus an eleventh locus lacking a selected locus were simulated. Recombination probabilities among loci within chromosomes were a function of distance between loci. A locus controlling mating system was located on a separate linkage group (chromosome *12*).

TABLE 1

Simulation conditions for modeling germplasm collections

Simulation condition	No. of demes per environment[a]	Mating system	Migration[b]
1	Equal	Outcrossing	Migration
2	Equal	Outcrossing	Isolation
3	Equal	Selfing	Migration
4	Equal	Selfing	Isolation
5	Uneven	Outcrossing	Migration
6	Uneven	Outcrossing	Isolation
7	Uneven	Selfing	Migration
8	Uneven	Selfing	Isolation

[a] Subdivided populations with "equal" distribution of demes had 6 demes per environment (population sizes of 2000, 1000, 1000, 500, 500 and 100). Subdivided populations with "uneven" distribution had 12 demes in environment I, four demes in environment II, and two demes in environment III. Deme sizes ranged from 100 to 3000 individuals, with a constant mean size of 1000.

[b] Migration was simulated by taking 2% of the individuals contributing to the next generation of each deme as migrants from other demes. Demes in the same environment contributed half of the migrants, while demes in the two remaining environments contributed the other half.

TABLE 2

Sampling strategies for core collections

Strategy	Sampling procedure	Genetic markers
R	Random sampling. Accessions sampled from the base collection without regard to group origin	None
C	Stratified random sampling[a], $\rho\hat{g} = 1/3$	None
P	Stratified random sampling[b], $\rho\hat{g} = \text{Size}\ (g)/[\Sigma\ \text{Size}(j)]$	None
L	Stratified random sampling, $\rho\hat{g} = \ln\ [\text{Size}\ (g)]/\{\Sigma\ \ln[\text{Size}(j)]\}$	None
M1–M20	Core collection constructed under two constraints[c]: (i) Include in each putative core collection at least one accession per environment. (ii) Maximize the marker allelic richness of the core collection.	One to 20 markers used to compute marker allelic richness of each putative core collection.

[a] Stratified sampling to build a core collection of size r involved random sampling of $r\rho_g$ accessions from each diversity group g.

[b] Size (g) denotes size in number of accessions of the group g. Σ denotes summation over the three groups. Index of summation j refers to group.

[c] The APPENDIX and text describe the optimization procedure.

whereas the remaining strategies involve different methods of stratified sampling in which $\rho_j * r$ accessions are randomly sampled from each diversity group j. The ρ_js reflect the weight of each group j in contributing to the core collection. The ρ_js can be identical, as in the constant (C) strategy or be based on some predictor of within group diversity. The proportional (P) and logarithmic (L) strategies use group size or natural logarithm of group size, respectively, as predictors of group diversity.

To employ marker locus data in the construction of core collections, we used the M strategy. While the M strategy is also a stratified sampling method, it is not based on random sampling of accessions within each diversity group. Rather, it chooses the specific combination of r accessions that maximizes the total allelic richness at available marker loci, subject to the constraint of including at least one accession per diversity group in the core collection (Table 2). In theory, this particular (optimal) combination of accessions could be discovered by examining every admissible combination of accessions and scoring these for marker allelic richness, but with a large base collection, inordinate computation time would be required. Instead, we implemented an algorithm that performed a series of heuristic searches thereby shortening the time required to find the optimal set of accessions (APPENDIX). The algorithm was tested by applying it to several smaller test cases (data from SCHOEN and BROWN 1993) where the set of accessions yielding maximum allelic richness was known by a prior complete search of all combinations of accessions. It succeeded in finding the correct combination of accessions in those cases.

To examine how the amount of marker locus information influences the effectiveness of the M strategy, we implemented it using 1, 5, 10, 15 or 20 loci. A random set of marker loci was chosen from the set of 90 neutral loci for each trial.

Once a core collection was created, its genetic diversity was assessed by counting the total number of alleles captured for two types of target loci: a random set of 10 selectively neutral loci ("neutral allelic richness"), different from marker loci used; and the 10 selected loci ("selected allelic richness"). As we were interested primarily in measuring total number of alleles captured rather than evenness of allelic frequencies, the allelic richness measures were not weighted by their frequencies. Allelic richness at loci under selection was used to compare the different sampling strategies. To assess the distributional properties of the sampling methods, 500 different core collections were assembled per strategy from each base collection.

RESULTS

Retention of allelic richness: Several broad scale patterns are apparent in the way that allelic richness was distributed at the level of population, base and core collections. These patterns can be best summarized once alleles are classified both as widespread (occurring in more than one deme) *vs.* localized (in only a single deme) and occurring at high frequency (frequency ≥ 0.1) *vs.* low frequency (frequency < 0.1) in each deme (MARSHALL and BROWN 1975). Thus at one extreme of this classification are alleles occurring in many demes and at high frequency in each deme, while at the other extreme are alleles occurring in only one deme and at low frequency. Because of their abundance, the former are unlikely to be lost during the creation of the base or core collections, regardless of which sampling method is used, whereas the latter are expected to be most vulnerable to loss (MARSHALL and BROWN 1975).

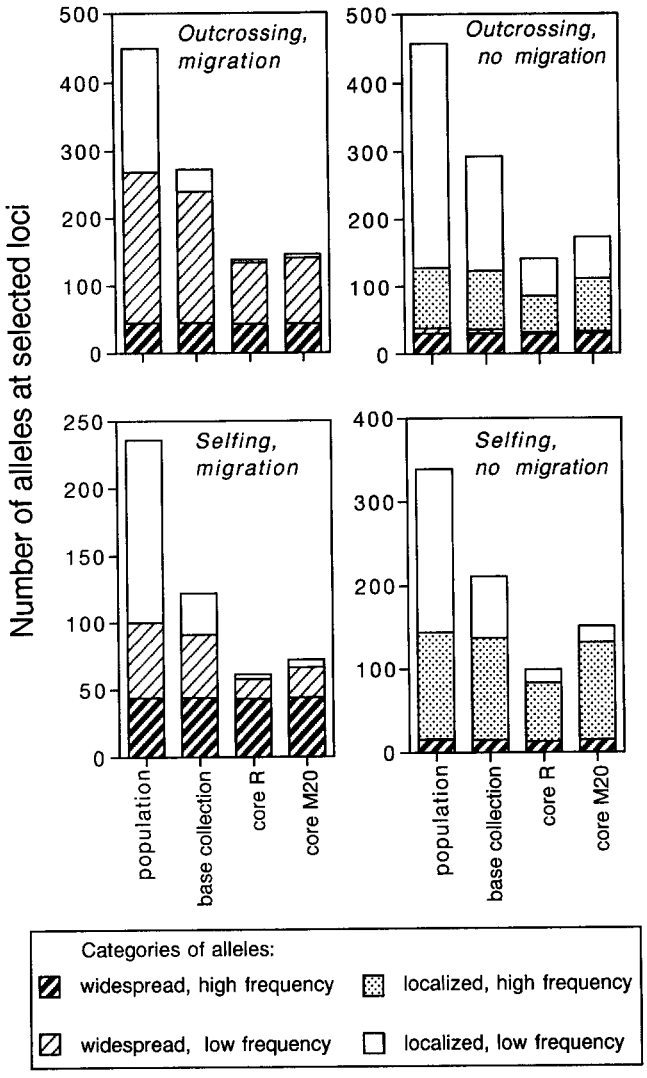

FIGURE 4.—Numbers and categories of alleles present at different stages in the simulation and sampling process. Categories of alleles are defined as follows: widespread, present in more than one deme; localized, present in only one deme; high frequency, at a frequency of ≥ 0.1 in at least one deme; low frequency, at a frequency of <0.1 in all demes. Results from simulations with equal number of demes in each environment are shown. Results for the R and M20 strategies are based on 500 independent trials. Standard errors not shown ($\ll 1$).

Figure 4 shows the numbers of the four different classes of alleles in the population, base collection, and core collection constructed under the R and M strategies (with 20 marker loci). Whereas the base and core collections contained approximately one-half and one-quarter of the total allelic richness of the subdivided population, respectively, most of what was lost during sampling were the localized, low-frequency alleles. Substantially more of the other classes of alleles were retained. Populations with different mating systems or migration rates differed systematically in terms of overall allelic richness and relative abundance of the four classes of alleles (Figure 4). For example, outcrossing

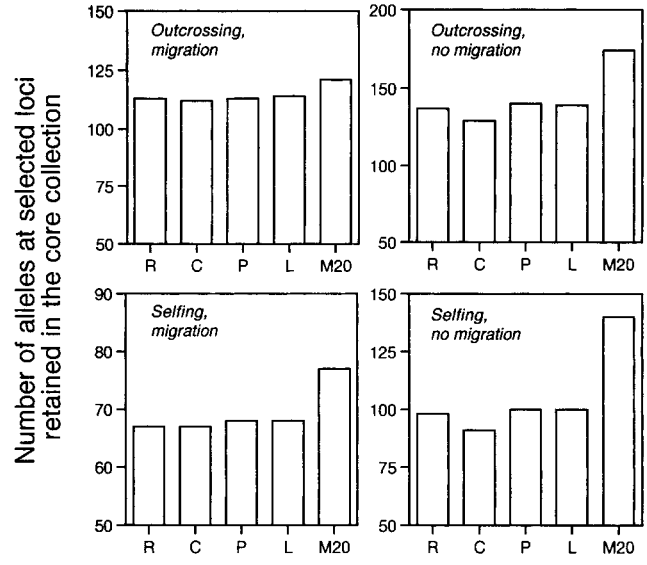

FIGURE 5.—Average allelic richness in core collections at loci under selection. Results for populations having uneven numbers of demes per environment are based on 500 replicate core collections per base collection. Sampling strategies are as follows: R, random; C, constant; P, proportional; L, logarithmically proportional; and M20, M strategy implemented with 20 markers. Height of each bar indicates the average allelic richness across the 500 independent replicates. Standard errors not shown ($\ll 1$).

populations generally contained higher allelic richness than selfing populations, consistent with their higher expected effective population size (POLLAK 1987). Migration among demes led to a large number of widespread, low-frequency alleles but few localized, high-frequency alleles. The opposite pattern was found when migration was not allowed (Figure 4).

Effectiveness of the different sampling strategies: The ranking of core collection allelic richness for the different types of sampling strategies was consistent across replicates under each specific combination of mating system, migration rate, and deme distribution. To simplify the presentation, therefore, results from only one replicate per combination are shown (Figures 5 and 6). Moreover, only results from variable number of demes per environments are shown, as they were qualitatively similar to those for equal number of demes. We focus on the number of selected alleles retained in the core collection. Results for retention of neutral alleles in the core collections were found to be similar to those seen for alleles at selected loci (data not shown).

For nonmarker based strategies, allele retention under the C strategy was slightly lower then for the other strategies, although the difference between it and the P and L strategies was small. Differences among the P, L and R strategies were also negligible. The only real contrast was between the allelic richness of core collections assembled using the nonmarker-based strategies and that observed using the M strategy. The increase in al-

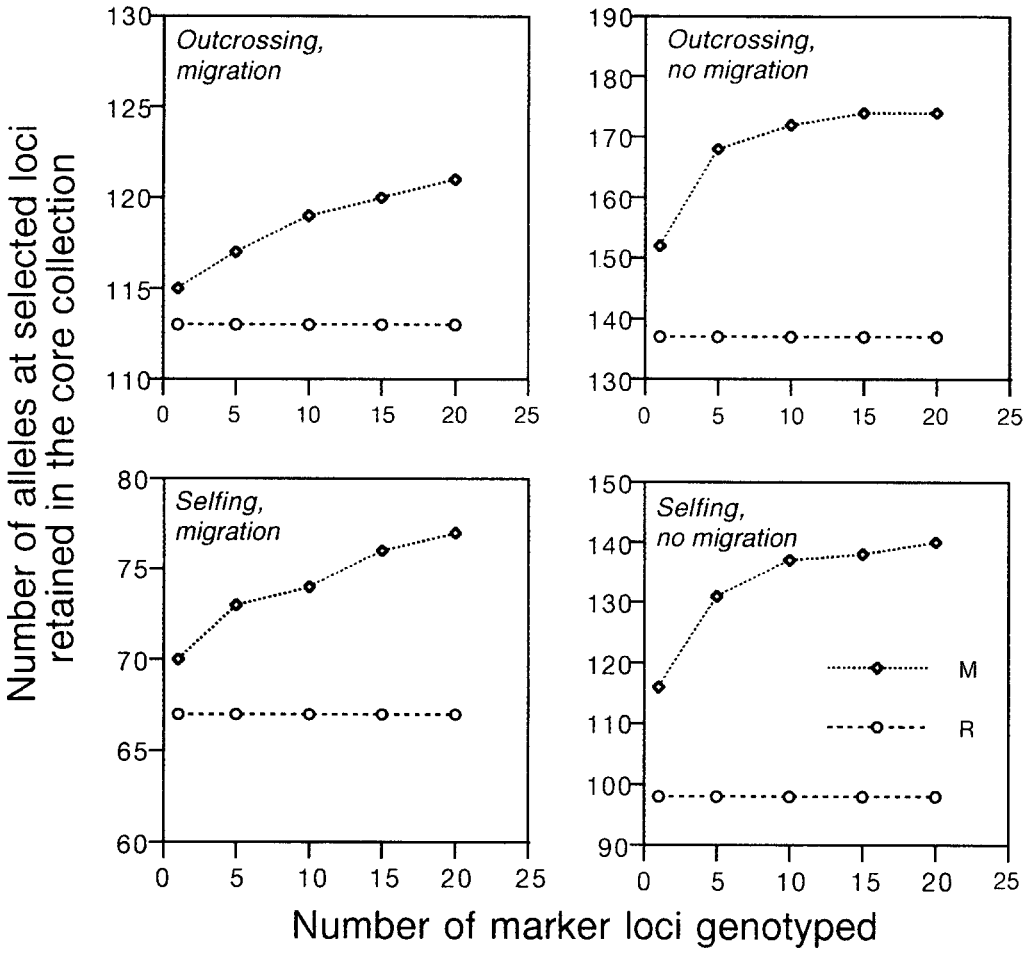

FIGURE 6.—Average allelic richness in core collections as a function of number of marker loci scored. Core collections assembled from populations with uneven distribution of demes. Performance of the R strategy is depicted for comparative purpose. Each data point is based on 500 independent trials. Standard errors not shown ($\ll 1$).

lelic richness under the M strategy was large and significant in all cases except under the combination of high outcrossing and migration. Elsewhere the M strategy led to conservation of markedly more alleles in the simulated core collections. The greater retention of alleles under the M strategy is due primarily to increased capture of the widespread, low frequency alleles and the localized, high frequency alleles (Figure 4). Migration and mating system had a dramatic effect on the relative performances of the M strategy. The overall impact of migration was to reduce the differences between the M and the other strategies, whereas selfing increased the efficiency of the M strategy. Selfing and absence of migration combined to reinforce the superiority of the M strategy over the other strategies (Figure 5).

Number of marker loci and temporal dynamics: Allele retention in core collections assembled under the M strategy increased markedly with the number of marker loci used (Figure 6). In simulations without migration, the performance of the M strategy plateaued with 10 marker loci, whereas with migration, the efficiency increased regularly with the number of maker loci used (Figure 6).

To study nonequilibrium behavior in the subdivided population, base and core collections were simulated and allelic richness was recorded after only 50, 200 and 2000 generations of evolution (as described above) following the initial creation of the subdivided population. We found that 2000 generations were required for the allelic richness of the population to achieve a pseudo-equilibrium (equivalent to what was seen after 5000 generations) (Figure 7). Moreover, the efficiency of the M strategy increased with the number of generations of genetic isolation ($\leq$2000 generations), in parallel with the appearance of many localized, high-frequency alleles (Figure 7).

DISCUSSION

Use of genetic marker data in genetic conservation: Information provided by surveying a large germplasm collection for molecular genetic variation could be put to use in a variety of ways to create a smaller, but genetically representative (core) collection. First, one might use clustering methods to classify accessions into groups using data from the markers, followed by stratified sampling within the groups (CROSSA *et al.* 1993; VAN HINTUM *et al.* 1995). The effectiveness of this approach remains to be examined.

Second, one could use marker data to assess neutral genetic diversity in a collection comprised of pre-existing groups formed on the basis of other criteria (*e.g.*,

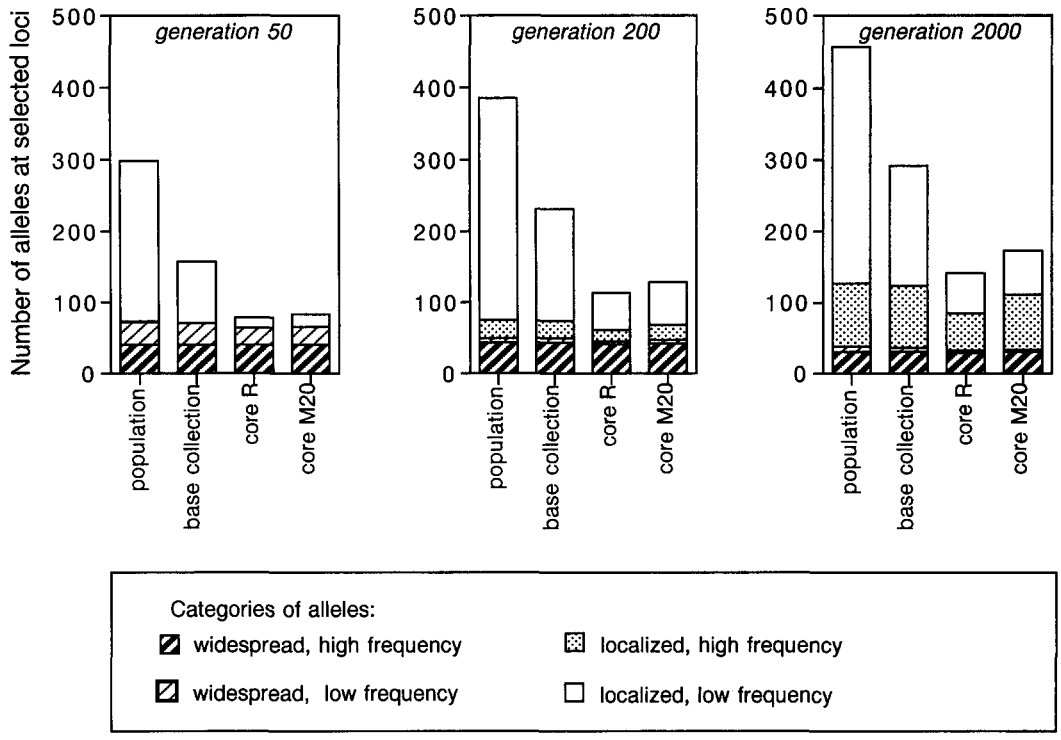

FIGURE 7.—Changes in allelic richness in populations or samples after 50, 200, or 2000 generations following the initial establishment of the original subdivided population. Results for a simulation involving outcrossing, no migration, and uneven distribution of demes.

ecogeographic data). The marker data could then be used to weight the relative contributions of each group to the core collection—*i.e.*, the "H strategy" of SCHOEN and BROWN (1993). Such a procedure is expected to perform well when groups are isolated and group size is a poor indicator of diversity (BROWN and SCHOEN 1994). For instance, populations that have recently passed through bottlenecks are expected to be largely depleted of low-frequency alleles (NEI *et al.* 1975), and so assessment of diversity at marker loci may be the best means of weighting the contributions of the groups to the core collection (SCHOEN and BROWN 1993). Since extinction and recolonization events were not part of the simulations, the H strategy was not examined in the present study.

Finally, with an already existing collection, marker data could be used to determine both the level of diversity of the groups comprising it (as in the H strategy), and the extent to which the accessions contain similar genetic variation ("redundancy"). This, in essence, is the thrust of the M strategy. According to our results, such an approach leads to significant increases in retention of allelic richness at both neutral and selected loci, whenever the collection shows any appreciable substructure. In fact, only when there is complete outcrossing and high migration rates did the M strategy not outperform the nonmarker-based strategies. This is likely due to the island model of migration used in our work—*i.e.*, immigrants came not only from demes with similar selective regimes but also from those with contrasting selection regimes, thereby counteracting adaptive differentiation. Together with high migration rates, F_{st} was near 0, meaning that neutral alleles within demes diverged only slightly more from a common ancestor compared with alleles from different demes (SLATKIN 1991). Moreover, the high level of migration and large deme size (average of 1000 individuals) counteracted the development of associations between neutral and selected loci. In general, the M strategy is never expected to perform more poorly than a random strategy (as verified above). For this to happen, there would have to be a negative correlation between diversity at marker and selected loci.

There is considerable controversy over the value of individual alleles in germplasm collections, especially localized, low-frequency alleles (BROWN 1978). These, however, are likely to be maintained by deleterious mutation-selection balance and, therefore, are of little interest in genetic conservation (MARSHALL and BROWN 1975). On the other hand, two classes of alleles—localized, high-frequency and widespread, low-frequency alleles—are of more interest, especially as the former may be a source of local adaptation. Interestingly, the M strategy led to more effective capture of such alleles. This improvement can be traced to redundancy in allelic composition of the accessions of the original base collection and correlations (shared coancestry) between among marker and target loci. The M strategy reduces the degree of redundancy in the final core collection. This alone would be unimportant, however,

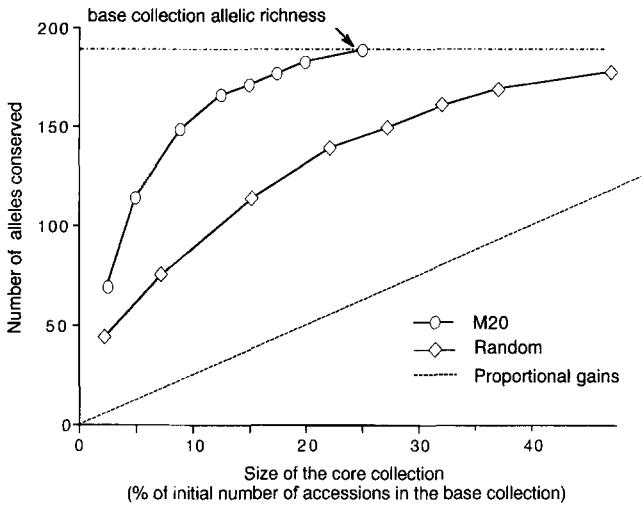

FIGURE 8.—Average allelic richness at marker loci in core collections as a function of collection size and sampling strategy. The horizontal dotted line indicates allelic richness of the base collection. Results from the R and M (using 20 marker loci) strategies are illustrated. Diagonal dotted line depicts the situation in which gains in allelic richness are proportional to core collection size (no redundancy). The simulation involved outcrossing, no migration, and equal number of demes in each environment. Each data point is based on 500 independent trials. Standard errors not shown ($\ll 1$).

were it not for the existence of correlations among loci, meaning that when redundancy is minimized for one class of observable variation (the marker loci) it will tend to be minimized for other, less easily assessed variation as well. Shared coancestry of alleles is likely due to subdivision in the population and base collection and the interaction between genetic drift and hitchhiking effects. Absence of gene flow, together with selfing (slowing the decay of linkage disequilibrium) may also have contributed to the maintenance of such correlations in the simulations.

Collection size and molecular markers: Hardy-Weinberg equilibrium and sampling with replacement have served as starting points to determine minimal sample sizes required for capture of at least one copy of an allele present at a given frequency (MARSHALL and BROWN 1975; CROSSA 1989). Using sampling theory for selectively neutral alleles (EWENS 1972), BROWN (1989a,b) showed that the number of alleles captured in a core collection was approximately proportional to the natural logarithm of its size, and sampling 10% of a population or collection leads to capture 60–70% of the alleles. An alternative approach for setting minimal collection size is to examine the actual empirical relationship between size of the collection and its genetic diversity, *e.g.*, as expressed by allelic richness at marker loci. Figure 8 shows this relationship for the simulations conducted in the present investigation. If a point of sharply diminishing returns (in terms of marker allelic richness) with further increase in the size of the core collection has not yet been reached at 10%, and sufficient resources are available, it may be justifiable to create a larger core collection. For instance in the present case, sampling 25% of the base collection (using the M strategy) allowed capture of all alleles at both neutral and selected loci (Figure 8). How many markers are required for such gains? In the simulations described here, a substantial increase ($\leq$40%) in allele retention at selected loci could be accomplished with as few as 5 to 10 marker loci. Preliminary results from a simulation of sampling a real "base collection" consisting of 120 divergent maize inbred lines (*e.g.*, Lancaster, Reid yellow dent, Minnesota 13, European flint, as well as other more exotic lines) (DE VIENNE *et al.* 1994) has shown that the M strategy implemented with 30—40 RFLPs as markers led to significant increases in allelic richness in the core collection (at other RFLP loci) (P. Dubreuil, A. Charcosset and T. Bataillon, unpublished results).

Our simulations corroborate earlier findings which showed that information from genetic markers leads to significant gains in conserved allelic richness (SCHOEN and BROWN 1993). Admittedly, the results presented here are based on a single pattern of selection and may not be general for all types of pattern of selection. Further work should investigate the impact of different modes of selection (*e.g.*, number of selective environments, patterns of selection) on genetic association between selected and neutral loci. But while allelic richness provides a straightforward measure of single locus variation in a collection, the question of how to best characterize quantitative genetic variation is more challenging and beyond the scope of the present effort. In the future, it would be of interest to explore the effectiveness of marker-based germplasm conservation strategies not only for quantitative genetic variation but also to examine how such strategies perform with intermediate to low rates of migration among demes and with partial self-fertilization. Explicit models of the dynamics of two locus linkage disequilibrium for neutral and selected loci in subdivided populations may also provide a better understanding of the population genetic mechanisms underlying the use of neutral markers in germplasm conservation.

We acknowledge CNUSC at Montpellier, France for providing access to their SP2 supercomputer. GUY DECOUX provided invaluable help in C language programming and access to several computer workstations. Many thanks to MARTIN MORGAN and KENT HOLSINGER for helpful and encouraging comments on earlier drafts of this paper. This work was supported by the International Council for Canadian Studies through a research fellowship to T.B. and an operating grant from the Natural Sciences and Engineering Research Council of Canada to D.J.S.

LITERATURE CITED

BROWN, A. H. D., 1978 Isozymes, plant population genetic structure and genetic conservation. Theor. Appl. Genet. **52:** 145–157.

BROWN, A. H. D., 1989a The case for core collections, pp. 136–156 in *The Use of Plant Genetic Resources,* edited by A. H. D. BROWN,

O. H. FRANKEL, D. R. MARSHALL and J. T. WILLIAMS. Cambridge University Press, NJ.

BROWN, A. H. D., 1989b Core collections: a practical approach to genetic resources management. Genome **31:** 318–324.

BROWN, A.H.D, 1995 The core collection at the crossroads, pp. 3–19 in *Core Collections of Plant Genetic Resources,* edited by T. HODGKIN, A. H. D. BROWN, T. J. L. VAN HINTUM and E. A. V. MORALES. John Wiley and Sons Ltd., Chichester, UK.

BROWN, A. H. D., and M. T. CLEGG, 1983 Isozymes assessment of plant genetic resources. Curr. Top. Biol. Med. Res. **11:** 285–295.

BROWN, A. H. D., and D. J. SCHOEN, 1994 Optimal sampling strategies for core collections of plant genetic resources, pp. 357–370 in *Conservation Genetics,* edited by V. LOESCHCKE, J. TOMIUK and S. K. JAIN. Birkhaüser Verlag, Basel, Switzerland.

CROSSA, J., 1989 Methodologies for estimating the sample size required for genetic conservation of outbreeding crops. Theor. Appl. Genet. **77:** 153–161.

CROSSA, J., C. M. HERNANDEZ, P. BRETTING, S. A. EBERHART and S. TABA, 1993 Statistical genetic consideration for maintaining germplasm collections. Theor. Appl. Genet. **86:** 673–678.

DAVID, J. L., Y. SAVY, and P. BRABANT, 1993 Outcrossing and selfing evolution in populations under directional selection. Heredity **71:** 642–651.

DE VIENNE, D., J. JOSSE, A. MAURICE, M. CAUSSE, A. LEONARDI *et al.,* 1994 Marquage et expression du génome chez le mais. Genet. Select. Evol. **26:** S21-S34.

ENDLER, J. A., 1986 *Natural Selection in the Wild.* Princeton University Press, Princeton, NJ.

EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87–112.

EWENS, W. J., 1979 *Mathematical Population Genetics.* Springer Verlag, Berlin.

FRANKEL, O.H., 1989 Principles and Strategies of evaluation, pp. 245–260 in *The Use of Plant Genetic Resources,* edited by A. H. D. BROWN, O. H. FRANKEL, D. R. MARSHALL and J. T. WILLIAMS. Cambridge University Press, NJ.

FRANKEL, O.H., and E. BENNETT, 1970 *Genetic Resources in Plants—Their Exploration and Conservation.* F.A. Davis, Philadelphia.

HENDERSON, S. T., and T. D. PETES, 1992 Instability of simple sequence DNA in *Saccharomyces cerevisae.* Mol. Cell. Biol. **12:** 2749–2757.

HOLSINGER, K. E., 1991 Conservation of genetic diversity in rare and endangered plants, pp. 626–633 in *The Unity of Evolutionary Biology: Proceedings of the Fourth International Congress of Systematic and Evolutionary Biology,* edited by E. C. DUDLEY. Dioscorides, Portland, OR.

HOSPITAL, F., and C. CHEVALET, 1996 Interactions of selection, linkage and drift in the dynamics of polygenic characters. Genet. Res. **67:** 77–87.

MARSHALL, D. R., and A. H. D. BROWN, 1975 Optimum sampling strategies in genetic conservation, pp. 53–80 in *Genetic Resources for Today and Tomorrow,* edited by O. H. FRANKEL and J. G. HAWKES. Cambridge University Press, Cambridge, UK.

MILLIGAN, B., J. LEEBENSMACK, and A.E. STRAND, 1994 Conservation genetics: beyond the maintenance of marker diversity. Mol. Ecol. **3:** 423–435.

NEI, M., T. MARUYAMA and R. CHAKRABORTY, 1975 The bottleneck effect and genetic variability in populations. Evolution **29:** 1–10.

POLLAK, E., 1987 On the theory of partially inbreeding finite populations: partial selfing. Genetics **117:** 353–60.

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.

SLATKIN, M., 1991 Inbreeding coefficients and coalescent times. Genet. Res. **58:** 167–175

VAN HINTUM, T. J. L., R. VON BOTHMER and D. L. VISSER, 1995 Sampling strategies for composing a core collection of cultivated barley (*Hordeum vulgare s lat*) collected in China. Hereditas **122:** 7–17.

WILSON, E. O., 1992 *The Diversity of Life.* Belknap Press of Harvard University Press, Cambridge, MA.

Communicating editor: A. H. D. BROWN

APPENDIX: IMPLEMENTATION OF THE M STRATEGY

Assume that a large collection has been stratified into b "diversity groups" and is composed of n accessions: $a_1, \ldots, a_n$. The entire set of accessions is characterized for k marker loci. Denote as $R(X)$ the allelic richness of an accession or group of accessions X scored at the k marker loci. To implement the M strategy, we wish to construct a core collection of size r ($r \geq b$), where for instance $r = 0.1*n$, subject to the constraint of including at least one accession originating from each of the b groups. Searching through all possible subsamples of size r is impractical when r and / or n exceed 50. Instead we employed an algorithm with the following steps:

Step 0. Within each group, the accession scoring the maximum $R(X)$ value is selected.

If $r > b$, then r-b accessions are sampled at random from the large collection, to form the current core collection (C).

Step 1. The r subsets ($S_1, \ldots, S_r$) that can be formed from C by removing one accession at a time are each considered, and their respective $R(S_1), \ldots, R(S_r)$ values are recorded. The subset S^* with the highest R value is retained.

Step 2. The accession a that brings the largest increase in new marker alleles into the core collection $C = (S^* + a)$ is chosen from the remaining accessions of the large collection.

Step 1 and 2 are repeated until $R(C)$ no longer changes. At that point, convergence is assumed and C is retained as the core collection.