*Proc. Natl. Acad. Sci. USA*
Vol. 90, pp. 10623–10627, November 1993
Population Biology

# Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers

**(genetic resources/core collections/population genetics/allozymes/sampling strategies)**

DANIEL J. SCHOEN* AND ANTHONY H. D. BROWN†

*Department of Biology, McGill University, 1205 Avenue Dr. Penfield, Montreal, PQ, Canada H3A 1B1; and †Division of Plant Industry, Commonwealth Scientific and Industrial Research Organization, GPO Box 1600, Canberra, Australian Capital Territory 2601, Australia

*Communicated by R. W. Allard, August 2, 1993*

ABSTRACT **Wild crop relatives are an important source of genetic variation for improving domesticated species. Given limited resources, methods for maximizing the genetic diversity of collections of wild relatives are needed to help spread protection over a larger number of populations and species. Simulations were conducted to investigate the optimal strategy of sampling materials from populations of wild relatives, with the objective of maximizing the number of alleles (allelic richness) in collections of fixed size. Two methods, based on assessing populations for variation at marker loci (e.g., allozymes, restriction fragment length polymorphisms), were developed and compared with several methods that are not dependent on markers. Marker-assisted methods yielded higher overall allelic richness in the simulated collections, and they were particularly effective in conserving geographically localized alleles, the class of alleles that is most subject to loss.**

Decisions in conservation biology may be based on demographic or genetic criteria or both (1). Demography often takes precedence when populations face immediate threats (2), but the long-term viability of species requires genetic variation (3). In domesticated species the conservation of single-locus variation is of special interest, as disease resistance and other economically important characteristics are often inherited in simple Mendelian fashion (4, 5). Genes introduced from the wild relatives of crops constitute an important source of single-locus variation for the improvement of domesticated species (6–8). The diversity of alleles at single loci (allelic richness) in wild relatives is particularly vulnerable to loss due to reduction in population size (9, 10). The problem is compounded by the fact that most crop relatives are found only in nature, and many such species and populations are increasingly threatened by habitat reduction (8). Moreover, there are many potentially useful populations of wild relatives, yet for practical purposes only a fraction of all such material can be afforded protection or maintenance in gene banks or in nature reserves. In addition, wild relatives are often geographically wide ranging, making it costly to collect representative samples of these materials. By maximizing genetic diversity in germ-plasm collections of fixed size, resources available for conservation of crop biodiversity can be allocated to a larger number of species.

This paper is concerned with how genetic markers (allozymes, restriction fragment length polymorphisms, etc.), which have been successfully employed in many other applied aspects of biology and medicine, can be used to help construct collections of wild crop relatives having maximal allelic richness. Because of their increasing importance in germ-plasm conservation we focus on the core collection (11, 12). The core collection consists of a limited number of accessions or populations selected from an existing larger collection (12). The core collection, a subsample of the whole collection, typically comprises about 10% of all available accessions and is intended to provide a set of genetically diverse material. The remaining accessions are not discarded, but priorities for evaluation of germ plasm begin with the core collection.

## MATERIALS AND METHODS

Two marker-assisted strategies for constructing core collections were developed. Each delineates how accessions are to be selected from a larger collection that has been divided into several broadly representative ecogeographic regions (stratified sampling).

For the first strategy, consider two regions and two polymorphic loci. With finite resources, the core collection comprises $n_1 + n_2$ (= constant) accessions from regions 1 and 2. If populations are isolated, an integral approximation of sampling theory for a model of selectively neutral loci whose allele frequencies are determined by mutation rate and genetic drift (infinite-alleles model) (12) gives the expected number of alleles retained in the core collection as

$$K \cong \theta_{11}\ln[\theta_{11} + n_1] + \theta_{21}\ln[\theta_{21} + n_1] + \theta_{12}\ln[\theta_{12} + n_2] + \theta_{22}\ln[\theta_{22} + n_2] - \sum_{ij} \theta_{ij}\ln\theta_{ij} + \text{constant}, \quad [1]$$

where $\theta_{ij}$ is an estimate of $\theta = 4N_e\nu$ ($N_e$ = effective population size, $\nu$ = the mutation rate) for the *i*th locus and *j*th region. The values of $n_1$ and $n_2$ that maximize $K$ ($\theta_{ij} \ll n_1$ and $n_2$) can be found by solving $dK/dn_1 \cong [\theta_{11}/n_1] + [\theta_{21}/n_1] - [\theta_{12}/n_2] - [\theta_{22}/n_2] = 0$. This gives $n_1/n_2 = (\theta_{11} + \theta_{21})/(\theta_{12} + \theta_{22})$; i.e., to achieve maximal core diversity, accessions from the two regions are sampled in proportion to the ratio of the sums of the $\theta_{ij}$. When generalized to $I$ loci ($i = 1, \ldots, I$) and $J$ regions ($j = 1, \ldots, J$) using Lagrange's multipliers, the optimal number of accessions to be sampled per region is

$$n_1{:}n_2{:}\cdots{:}n_J = \left(\sum_i \theta_{i1}\right){:}\left(\sum_i \theta_{i2}\right){:}\cdots{:}\left(\sum_i \theta_{ij}\right), \quad [2]$$

where $\Sigma_j n_j$ = constant. The approach is referred to as the "H strategy" (after Nei's index of gene diversity—see below).

In the second marker-assisted conservation strategy, populations are assumed to exchange genes. It is of interest, therefore, to select accessions that each contain many alleles, while minimizing overlap in allelic composition between accessions. Consequently, unlike the H strategy, this approach pinpoints the individual accessions to be sampled from each geographic region. It relies on correlation of diversity and identity among separate loci. Correlation of diversity is expected when populations differ in the magnitude or recency of past bottlenecks or the strength of directional selection, leading to overall genome-wide variation

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

among populations in allelic richness. In such cases, levels of allelic richness detected at marker loci may be useful as a guide to those at other loci. Correlation of identity arises because accessions that have a recently shared evolutionary ancestry will have many alleles in common at separate loci (13). Thus a subset of all the available accessions which are well differentiated from one another at the marker loci will likewise be well differentiated from one another at other (target) loci of interest to genetic conservation. Let $A_{aj}$ denote one of many possible subsets of all accessions (i.e., $A_{aj}$ is a candidate for the core collection and is simply the set of each accession $a$ selected from each region $j$). Let $Q_i\{A_{aj}\}$ = the probability that marker allele $i$ is *not* retained when $A_{aj}$ is selected as the core collection. The problem of constructing a core collection with high allelic richness can be stated in linear programming notation as follows:

Minimize: $\sum_i Q_i\{A_{aj}\}$

Subject to: (*i*) a core collection of constant size (e.g., 10% of all populations or accession)
(*ii*) at least one population or accession per region (i.e., $n_j \geq 1$).

The method is referred to as the M strategy (marker allele richness). With $\Sigma_j n_j \geq 40$ accessions there are often more than $10^7$ ways of selecting $n_j \geq 1$ accessions from each of $J$ different regions. A computer program (MSEARCH) was written to search through all possible $A_{aj}$ and identify the one(s) that minimized $\Sigma_i Q_i\{A_{aj}\}$. The program is available from D.J.S.

To test the effectiveness of the H and M strategies, core collections of wild relatives of several grain, vegetable, and fiber crops were simulated by computer. The simulations were conducted with published allozyme and geographical data (Table 1). Allozyme loci were divided randomly into two halves—the first half served as markers and were used only to assist in sampling by the H and M strategies, while the remaining half were treated as targets of conservation whose allelic richness in the core was assessed for each strategy.

Simulated core collections were assembled by sampling populations or accessions from separate ecogeographic regions in each crop relative according to the strategy at hand. When sampled, the accession (and all its alleles) were included as entries in the core collection. For the H strategy, individual estimates of $\theta_{ij}$ were obtained as $\theta_{ij} = h_{ij}/(1 - h_{ij})$, where $h_{ij}$ is Nei's gene diversity index for the $i$th locus and $j$th region (22). The expected number of target alleles retained in the core was calculated by sampling accessions without replacement and was obtained as the mean over 300 simulation trials.

For comparison with the H and M strategies, simulated core collections were also constructed by using a number of methods that do not rely on markers. These other methods

Table 1. Taxa and loci used for constructing and testing simulated core collections of wild crop relatives

| | | | No. of accessions | | No. of loci or alleles | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Species* | Origin | Regions for stratified sampling | Total | Core collection‡ | Marker loci (allozymes) | Marker alleles | Target loci (allozymes) | Target alleles | Correlation of gene diversity† | Data source (ref.) |
| *Hordeum spontaneum* | Israel | 1. Inland—mesic<br>2. Inland—xeric<br>3. Coastal | 28 | 6 | 15 | 50 | 10 | 51 | 0.05 | 14 |
| *Sorghum bicolor* ssp. *arundinaceum* | Africa | 1. South<br>2. East<br>3. West | 68 | 6 | 4 | 54 | 11 | 42 | 0.07 | 15 |
| *Zea mays* ssp. *parviglumis/ mexicana* | Mexico | 1. North<br>2. South | 37 | 4 | 10 | 81 | 12 | 73 | 0.08 | 16 |
| *Lycopersicon pimpinellifolium* | Peru | 1. Piara, Lambayeque<br>2. La Libertad, Ancash<br>3. Lima | 35 | 6 | 5 | 25 | 5 | 12 | 0.20 | 17 |
| *Solanum pennellii* | Peru | 1. Lima<br>2. Ica | 21 | 4 | 8 | 38 | 7 | 37 | 0.11 | 18 |
| *Capsicum annum/frutescens* | Mexico | 1. Campeche, Tabasco to Jalisco, Michoacan<br>2. Nuevo Leon, Tamaulipas<br>3. Sonora | 66 | 6 | 10 | 34 | 9 | 27 | 0.06 | 19 |
| *Phaseolus vulgaris* | Central America<br>South America | 1. Mexico, Guatemala, Costa Rica<br>2. Colombia, Peru<br>3. Argentina | 70 | 6 | 4 | 11 | 4 | 12 | —§ | 20 |
| *Solanum berthaultii/tarijense* | Bolivia<br>Argentina | 1. North<br>2. South | 29 | 4 | 5 | 13 | 5 | 14 | 0 | ¶ |
| *Gossypium davidsonii* | Baja California, Mexico | 1. Cabo San Lucas<br>2. La Paz | 13 | 4 | 7 | 13 | 6 | 13 | −0.11 | 22 |

*In several instances, two closely related taxa were pooled and treated as one.

†The correlation of gene diversity, $R$, was calculated as $R = \Sigma_i \Sigma_{k>i} \{\mathrm{cov}(h_{ia}, h_{ka})/[\mathrm{var}(h_{ia})\ \mathrm{var}(h_{ka})]^{1/2}\}/\{I(I - 1)/2\}$, where $h_{ia}$ is Nei's gene diversity index for the $i$th locus and the $a$th accession. $R$ provides a measure of association between single-locus gene diversities and is positive when populations that exhibit high (or low) diversity at one locus also tend to exhibit high (or low) diversity at other loci. Variation among taxa in $R$ for marker loci reflects that seen in all loci as shown above.

‡Larger core collections (>10% of total) were simulated in some cases to allow a clearer contrast between the different core collection strategies.

§Data available on allele numbers only.

¶D. M. Spooner and D. S. Douches, personal communication.

included the sampling of a constant number of accessions per region for entry into the core collection (C strategy), sampling in proportion to the number of accessions available per region (P strategy), sampling in proportion to the logarithm of the number of accessions available per region (L strategy), and simple random (not stratified) sampling of accessions (R strategy) (12). The expected number of target alleles retained in these simulated core collections was calculated as above.

## RESULTS

Mean ranking of core target allele retention for the six strategies in the nine test cases (highest rank = 1, lowest = 6) was M (2.1) > H (3.0) > P (3.1) > L (3.7) > C (4.1) > R (4.7) (Fig. 1). The two marker gene-assisted strategies thus yielded core collections with the highest overall allelic richness. Target allele retention was maximized under the M strategy in all but two taxa. The two exceptions, *Solanum berthaultii/tarijense* and *Gossypium davidsonii*, were the only taxa that did not exhibit positive correlation of gene diversity, probably due to sampling error associated with the small number of available accessions and loci (Table 1). Such exceptions, however, can be spotted in advance by calculating the correlation of diversity for the marker locus fraction (Table 1). With these species excluded, the mean ranking of core allele retention was M (1.0) > H (2.9) > P (3.5) > L (3.8) > C (4.4) > R (5.0). Moreover, when marker-assisted methods were used, there was a significant improvement in the retention of alleles found in only one or a few accessions (Fig. 2). Localized alleles such as these are the most likely ones to be lost from the core collection. Localized alleles may also represent sources of resistance to local pathogen races or provide adaptation to specific environmental conditions (23).

## DISCUSSION

An effective genetic conservation strategy should maximize the retention of genetic variation associated with long-term species survival, yet it is not feasible to assess directly the level and distributional properties of such variation (23) let alone predict which traits might become important in future environments. While data on marker locus diversity are straightforward to obtain, there has been some concern as to whether the distribution patterns of marker and target genes are sufficiently similar to justify using information gained from markers (24). Several points are relevant to this question. First, the H and M strategies are not guided exclusively by marker data. The initial step of stratification by geographical region or habitat is an ecological decision, made on the basis of knowledge about the species' range and environmental amplitude. This step is likely to capture a portion of the existing adaptive variation. The use of markers following stratification might then best be viewed as a means to achieve an appropriate allocation of sampling effort within and among regions. Second, at least some of the variation that is potentially useful in future environments may at present be selectively neutral, and consequently its distributional properties

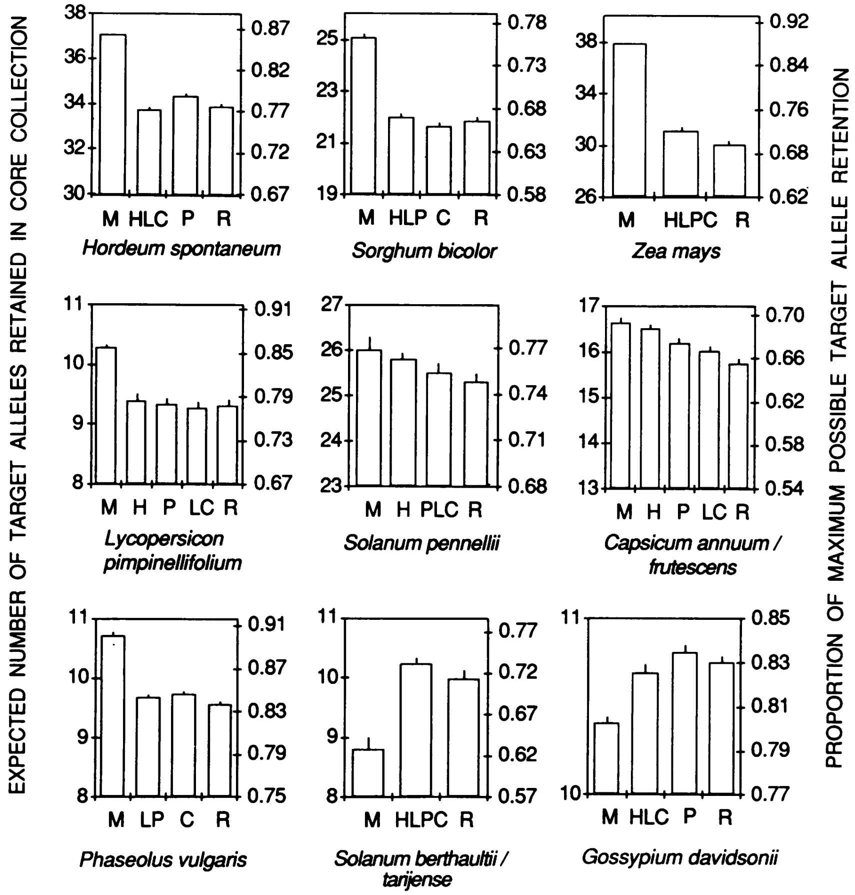
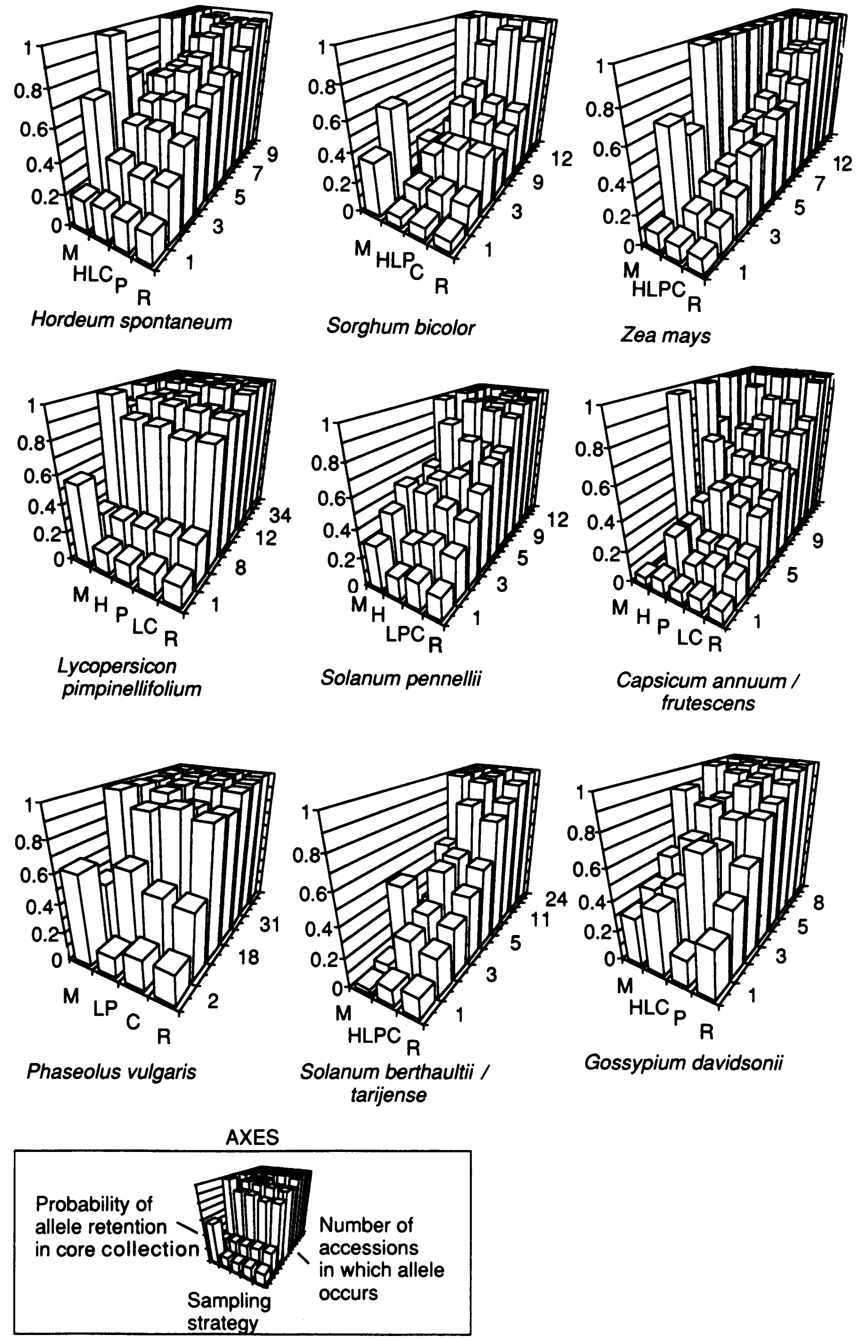


FIG. 1. Expected target allele retention in simulated core collections and proportion of maximal possible allele retention as a function of core collection strategy (SE indicated by line above bar). Core collection strategies other than H and M are as follows: sampling of a constant number of accessions per region for entry into the core collection (C strategy), sampling in proportion to the number of accessions available per region (P strategy), sampling in proportion to the logarithm of the number of accessions available per region (L strategy), and simple random (not stratified) sampling of accessions (R strategy) (12). The H strategy was not examined in *Phaseolus vulgaris* due to absence of allele frequency data. Values are for core collections with the sizes indicated in Table 1.



*Hordeum spontaneum* *Sorghum bicolor* *Zea mays*

*Lycopersicon pimpinellifolium* *Solanum pennellii* *Capsicum annuum / frutescens*

*Phaseolus vulgaris* *Solanum berthaultii / tarijense* *Gossypium davidsonii*

FIG. 2. Allele retention in simulated core collections as a function of sampling strategy and frequency of occurrence of target allele among all accessions. The probability of allele retention was determined as the proportion of all simulated core collections (300 trials) containing the allele in question. Simulation methods and core collection strategies are as described under Fig. 1.

may be approximated by recourse to markers. Third, a marker-assisted approach may be especially appropriate in inbreeding species, which comprise a large proportion of crop relatives (6). In inbreeders, variation in the whole genome is often correlated due to a slowdown in the decay of linkage disequilibrium (25, 26). Thus, maximizing the allelic richness at marker loci in inbreeders can lead to increased allelic richness at other loci. Moreover, inbreeders exhibit much more variation in allelic richness among populations than do outbreeders (21), and this is precisely the situation in which assembling a diverse core collection may be most difficult in the absence of genetic data.

We thank J. Doebley, D. Douches, K. Ritland, D. Spooner, and C. Stuber for sharing information on crop relatives. J. Burdon, P. Gepts, K. Holsinger, M. Johnston, P. Larkin, and G. Lawrence read the manuscript and provided helpful comments. This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

1. Soulé, M. E. (1987) *Viable Populations for Conservation* (Cambridge Univ. Press, Cambridge).
2. Lande, R. (1988) *Science* **241,** 1455–1460.
3. Frankel, O. H. (1970) *Proc. Linn. Soc. N. S. W.* **95,** 158–169.
4. Burdon, J. J. (1987) *Diseases and Plant Population Biology* (Cambridge Univ. Press, Cambridge).
5. Hilu, K. W. (1983) *Evol. Biol.* **16,** 97–128.
6. Allard, R. W. (1960) *Principles of Plant Breeding* (Wiley, New York).
7. Chapman, C. G. D. (1989) in *The Use of Plant Genetic Resources*, eds. Brown, A. H. D., Frankel, O. H., Marshall,

D. R. & Williams, J. T. (Cambridge Univ. Press, Cambridge), pp. 136–156.

8. Hoyt, E. (1992) *Conserving the Wild Relatives of Crops* (Int. Board of Plant Genetic Resources, Rome).
9. Nei, M., Maruyama, T. & Chakraborty, R. (1975) *Evolution* **29**, 1–10.
10. Sirkkomaa, S. (1983) *Hereditas* **99**, 11–20.
11. Frankel, O. H. (1984) in *Genetic Manipulation: Impact on Man and Society*, eds. Arber, W., Llimensee, K., Peacock, W. J. & Starlinger, P. (Cambridge Univ. Press, Cambridge), pp. 161–170.
12. Brown, A. H. D. (1989) *Genome* **31**, 818–824.
13. Wright, S. (1969) *Evolution and the Genetics of Populations* (Univ. of Chicago Press, Chicago), Vol. 2.
14. Nevo, E., Zohary, D., Brown, A. H. D. & Haber, M. (1979) *Evolution* **33**, 815–833.
15. Morden, C. W., Doebley, J. F. & Schertz, K. F. (1990) *Theor. Appl. Genet.* **80**, 296–304.
16. Doebley, J. F., Goodman, M. M. & Stuber, C. W. (1984) *Syst. Bot.* **9**, 203–218.
17. Rick, C. M., Fobes, J. F. & Holle, M. (1977) *Plant Syst. Evol.* **127**, 139–170.
18. Rick, C. M. & Tanksley, S. D. (1981) *Plant Syst. Evol.* **139**, 11–45.
19. Loaiza-Figueroa, F., Ritland, K., Laborde Cancino, J. A. & Tanksley, S. D. (1989) *Plant Syst. Evol.* **165**, 159–188.
20. Koenig, R. & Gepts, P. (1989) *Theor. Appl. Genet.* **78**, 809–817.
21. Schoen, D. J. & Brown, A. H. D. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 4494–4497.
22. Wendel, J. F. & Percival, A. E. (1990) *Plant Syst. Evol.* **171**, 99–115.
23. Nei, M. (1973) *Proc. Natl. Acad. Sci. USA* **70**, 3321–3324.
24. Marshall, D. R. & Brown, A. H. D. (1975) in *Crop Genetic Resources for Today and Tomorrow*, eds. Frankel, O. H. & Hawkes, J. G. (Cambridge Univ. Press, Cambridge), pp. 53–80.
25. Holsinger, K. E. (1991) in *The Unity of Evolutionary Biology: The Proceedings of the Fourth International Congress of Systematic and Evolutionary Biology*, ed. Dudley, E. C. (Dioscorides, Portland, OR), pp. 626–633.
26. Hedrick, P. W., Jain, S. K. & Holden, L. R. (1978) *Evol. Biol.* **11**, 101–184.