Theor Appl Genet (2012) 124:1155–1171
DOI 10.1007/s00122-011-1776-4

ORIGINAL PAPER

Retention of agronomically important variation in germplasm core collections: implications for allele mining

Patrick A. Reeves · Lee W. Panella · Christopher M. Richards

Received: 19 July 2011 / Accepted: 15 December 2011 / Published online: 7 January 2012

Abstract The primary targets of allele mining efforts are loci of agronomic importance. Agronomic loci typically exhibit patterns of allelic diversity that are consistent with a history of natural or artificial selection. Natural or artificial selection causes the distribution of genetic diversity at such loci to deviate substantially from the pattern found at neutral loci. The germplasm utilized for allele mining should contain maximum allelic variation at loci of interest, in the smallest possible number of samples. We show that the popular core collection assembly procedure "M" (*marker* allele richness), which leverages variation at neutral loci, performs worse than random assembly for retaining variation at a locus of agronomic importance in sugar beet (*Beta vulgaris* L. subsp. *vulgaris*) that is under selection. We present a corrected procedure ("M+") that outperforms M. An extensive coalescent simulation was performed to demonstrate more generally the retention of neutral versus selected allelic variation in core subsets assembled with M+. A negative correlation in level of allelic diversity between neutral and selected loci was observed in 42% of simulated data sets. When core collection assembly is guided by neutral marker loci, as is the current common practice, enhanced allelic variation at agronomically important loci should not necessarily be expected.

Communicated by A. Graner.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-011-1776-4) contains supplementary material, which is available to authorized users.

P. A. Reeves (✉) · C. M. Richards
National Center for Genetic Resources Preservation, United States Department of Agriculture, Agricultural Research Service, 1111 South Mason Street, Fort Collins, CO 80521, USA
e-mail: pat.reeves@ars.usda.gov

L. W. Panella
Northern Plains Area Sugarbeet Research Unit, United States Department of Agriculture, Agricultural Research Service, 1701 Centre Ave, Fort Collins, CO 80521, USA

Introduction

Adaptations are the currency of biodiversity. Genetically based adaptive diversity, which has arisen via the process of natural selection, is the principal object of interest in fields attempting to understand, preserve, or utilize biodiversity. Adaptations, and the genes that underlie them, are a valuable resource for evolutionary biologists, conservation managers, and agricultural researchers. Adaptive variation, in addition to its importance for ensuring the success of the species in which it originated, has great intellectual, cultural, and economic value to humankind.

Gene banks operate at a unique nexus between evolutionary biology, conservation genetics, and crop science, where description, preservation, and utilization of adaptive diversity play equally important roles (Schoen and Brown 2001; Börner 2006; Walters et al. 2008). By collecting biodiversity from nature, establishing long-term storage, and organizing and distributing genetic variation to users, gene banks broker the benefits of adaptive biodiversity from nature to human society. Unfortunately, efficient extraction and exploitation of the adaptive variation and valuable traits maintained in gene banks have yet to be fully achieved, though it remains a high priority of gene bank managers (Hoisington et al. 1999; Richards 2004). Traditional methods, which screen large, heterogeneous collections for phenotypic variation in agricultural traits, are not only logistically challenging but they may overlook

valuable genotypic variation concealed by epistasis in non-elite genetic backgrounds (Tanksley and McCouch 1997).

Allele mining offers the prospect for expedited recovery of useful adaptations from gene banks. Allele mining experiments seek to identify naturally occurring allelic variants at loci of agronomic importance, i.e. those genes that affect crop characteristics and performance. Agronomic loci have been identified using a variety of approaches including mutant screens (Johal and Briggs 1992; Whitham et al. 1994; Bishop et al. 1996), QTL analysis (Backes et al. 1995; Xiao et al. 1996; Bernacchi et al. 1998), association mapping (González-Martínez et al. 2007; Crossa et al. 2007), and genomewide surveys for the signature of artificial selection (Vigouroux et al. 2002; Casa et al. 2005; Yamasaki et al. 2005; Chapman et al. 2008). Novel alleles recovered at loci of agronomic importance can be integrated into crop breeding programs using conventional or molecular approaches, and might be utilized to combat disease (Caicedo 2008; Kaur et al. 2008; Wang et al. 2008; Bhullar et al. 2009, 2010), to promote yield increases, to produce better storage and nutritional properties, or to improve stress tolerance (Latha et al. 2004).

The success of allele mining operations is dependent on the availability of diverse germplasm collections (Kumar et al. 2010). The majority of allelic variation at any given locus is predicted to occur in the wild relatives of a crop, and not the crop itself, due to the inevitable loss of variation at the domestication bottleneck (Tenaillon et al. 2004; Hyten et al. 2006; Zhu et al. 2007). Thus allele mining efforts will increasingly focus on wild material to identify useful new alleles not already present in the crop gene pool (Tanksley and McCouch 1997; Gur and Zamir 2004; Johal et al. 2008; Prada 2009). The curation of wild germplasm collections is complicated by this wealth of diversity. Organizing natural genetic variation for efficient characterization and exploitation, without a priori knowledge of the phenotype that will be targeted for improvement, is a major challenge for contemporary gene banking.

The core collection, a representative subset of the complete collection that has been optimized to contain maximal diversity in a minimal number of accessions, has been the primary solution proposed for facilitating the utilization of diverse germplasm collections (Frankel 1984; Brown 1989). Core collections are designed to streamline the integration of new, useful alleles into conventional breeding programs by reducing the number of accessions necessary in experimental crosses or phenotypic screening studies while maintaining, to the maximum extent possible, allelic diversity at loci controlling traits of interest. The potential improvement in screening efficiency offered by the core collection concept to conventional breeding is equally applicable to modern allele mining efforts. But, in order to identify subsets of the collection optimized for discovering the alleles necessary to solve future challenges, it will be necessary to predict the level of variation present at an undiscovered target locus affecting a thus far undetermined phenotype using simple-to-obtain information, such as ecogeographical attributes of sampling localities or molecular marker variation at easily assayed reference loci.

A number of optimization procedures for assembling core collections have been proposed. Most methods use some form of a priori splitting, or stratification, of the accessions into "diversity groups" that typically reflect ecogeographical differences among the original sampling localities (Brown 1989). Following an initial stratification step, core sets can be assembled by selecting a *constant* number of accessions at random from each diversity group (the "C" procedure), by choosing accessions at random in *proportion* to the size of the diversity group ("P" procedure) or in proportion to the *logarithm* of the size of the diversity group ("L"), or by choosing accessions at random in proportion to an estimate of *heterozygosity* within diversity groups ("H") (Schoen and Brown 1993). Other methods for core assembly include selecting accessions in proportion to the mean genetic *distance* between individuals within diversity groups ("D", Franco et al. 2005; "genetic distance sampling", Jansen and van Hintum 2007), and choosing accessions such that the total allelic diversity at a set of neutral reference loci is maximized in the resulting subset, provided that at least one accession from each diversity group is included ("M", for *marker* allelic richness; Schoen and Brown 1993). In practice, both genetic distance sampling and the M procedure do not require initial stratification, an advantage in our opinion because divisions based on ecogeographic region might be arbitrary with respect to the distribution of allelic diversity at loci of agronomic importance. Of all procedures, M has been applied most frequently, in part because of its demonstrated efficiency for retaining maximal allelic diversity at reference loci in a small subset of accessions (e.g. McKhann et al. 2004; Balfourier et al. 2007; Escribano et al. 2008; Le Cunff et al. 2008).

Under the M procedure, the selection of accessions is guided by patterns of variation at neutral reference loci, with the final subset consisting of those accessions that, collectively, contain the highest possible number of distinct alleles at the reference loci. Using simulated data, M has been shown to be effective for assembling subsets that, additionally, retain elevated levels of diversity at target loci of interest, i.e. loci that have not been used to guide assembly (Bataillon et al. 1996). The finding that the number of allelic variants retained at unlinked target loci is also maximized by applying M suggests that allelic diversity at one locus within an accession (or population) is correlated with allelic diversity at other loci within that accession. The results from the simulations of Bataillon

et al. (1996) thus lead to the hypothesis that allelic diversity is not simply a property of an individual locus, but is a property of populations as well—those populations with high allelic diversity at one set of loci (in this case, the reference loci) tend to have high allelic diversity at other, independent loci (the target loci).

Patterns of variation at neutral loci, including properties such as total allelic diversity, reflect population demography and shared ancestry among members. Because such historical attributes are common to all members of a population, allelic diversity is expected to be correlated across unlinked neutral loci. For loci under selection, however, the pattern may be very different and run counter to that predicted by demographic history and common ancestry (Reed and Frankham 2001; McKay and Latta 2002; Charlesworth et al. 2003). Indeed, it is precisely these differences that are leveraged in genomewide scans for the signature of selection (Nielsen 2005; Wright and Gaut 2005; Walsh 2008). These scans identify genes of agronomic importance as those genes inferred to have been under positive (directional) selection during domestication based upon a deviation in allele frequency spectrum from neutral expectations (Vigouroux et al. 2002; Casa et al. 2005; Yamasaki et al. 2005).

To produce subsets of germplasm collections for efficient allele mining at agronomic loci using the M procedure, one must assume that allele counts at neutral reference loci are predictive of allele counts at selected target loci (i.e. target loci with a history of natural or artificial selection). This assumption is questionable. Allelic diversity is primarily controlled by population size (Frankham 1996). For neutral loci, allelic variation is driven by the balance between mutation and extinction such that the larger the population, the greater the number of segregating neutral alleles ($n = 4N_e\mu + 1$ at equilibrium, where n is the number of alleles, N_e is the effective population size, and μ is the mutation rate) (Kimura and Crow 1964). A converse relationship holds for selected loci. Directional selection, which winnows segregating variation, is most efficient with large population sizes because allele frequencies are less affected by genetic drift. As population size decreases, alleles under selection become, in effect, neutral in terms of their expected frequency distribution. This occurs when N_e drops below $1/2s$, where s is the selection coefficient (Wright 1931). Thus, at selected loci, larger populations do not necessarily have greater numbers of segregating alleles. As stated most plainly by Frankham (1996), "The relationship between genetic variation and population size should be strongest for neutral genetic markers and poorest for the most strongly selected markers…" (p. 1502). Hence we should not expect allele counts at neutral and selected loci to be correlated within any given population. Nevertheless, the simulation study of Bataillon et al. (1996) states that "Results for retention of neutral alleles in the core collections were found to be similar to those seen for alleles at selected loci…" (p. 413): in other words, that the M procedure performed similarly regardless of whether target loci were neutral or selected. This finding seems counter to theory.

In this study, we use both empirical and simulated data to examine the potential of the M procedure to improve the efficiency of allele mining at loci of agronomic interest. By doing so, we re-examine the suggestion that allelic diversity is correlated between neutral and selected loci. We consider molecular polymorphism data from germplasm accessions derived from native populations of the sea beet, *Beta vulgaris* subsp. *maritima*, the wild progenitor of the sugar beet, collected along the Mediterranean and Atlantic coasts of France. The target locus, *BvFL1*, is a homolog of the *Arabidopsis* flowering time gene *FLC* and a locus of agronomic importance in sugar beet (Reeves et al. 2007). A coalescent simulation of neutral and selected loci under a broad range of population structures and demographic histories is performed to provide generality to conclusions.

Materials and methods

Acquisition of molecular polymorphism data

DNA was sampled from 28 *Beta vulgaris* subsp. *maritima* germplasm accessions preserved in the U.S. National Plant Germplasm System (Table 1). Accessions contain progeny from collections of wild populations made along the Atlantic and Mediterranean coasts of France, including five populations from Corsica (Fig. 1a). Two hundred eighty-four individuals were sampled, with 8 or 12 individuals representing each population. Twelve SSR (simple sequence repeat) loci were genotyped for all individuals as described previously (Viard et al. 2002; Richards et al. 2004; McGrath et al. 2007). These loci were treated as reference loci to guide subset assembly. The likelihood ratio test for linkage among loci implemented in FSTAT (Goudet 2001) identified two pairs of linked loci. For each pair, the locus with fewer alleles was eliminated, resulting in a reference data set containing codominant genotypes at ten unlinked SSR loci. Allelic richness was calculated using the rarefaction method of El Mousadik and Petit (1996) in FSTAT.

Two SSR loci tightly linked to the gene *BvFL1* were used as target loci to test the potential for anonymous reference loci to predict variation at agronomic loci. The loci were identified within the genomic *BvFL1* sequence (EF036526) and genotyped for all individuals. Locus *509* is located ~24 Kbp upstream from the *BvFL1* start codon; locus

Table 1 *Beta vulgaris* subsp. *maritima* germplasm accessions sampled

PI	*n*	Latitude	Longitude
504266	12	41.3889	9.1656
504269	8	41.5167	9.2167
504279	12	41.9161	8.7392
504277	12	42.6328	8.9422
504273	8	42.7022	9.4508
540562	12	43.4667	4.3333
540557	12	44.0978	0.2772
540578	12	45.3000	−0.7833
540582	8	45.8000	−1.1500
540592	12	46.2442	−1.5611
540595	8	46.3000	−1.0167
540599	8	46.6333	−1.8667
540602	12	47.0167	−1.9667
540606	12	47.2333	−2.1500
540609	8	47.6667	−3.1667
540692	8	47.7664	−3.5486
540613	12	47.9333	−4.3908
540618	8	48.5167	−4.7500
540637	8	48.6167	−2.0333
540640	12	48.6333	−1.4833
540641	12	48.6500	−1.4000
540690	8	48.7828	−3.0575
540645	12	49.0000	−1.5500
540656	8	49.2833	−0.1333
540647	12	49.3333	−1.7000
540651	8	49.5833	−1.2667
540661	12	50.0167	1.3333
540665	8	50.7667	1.6167

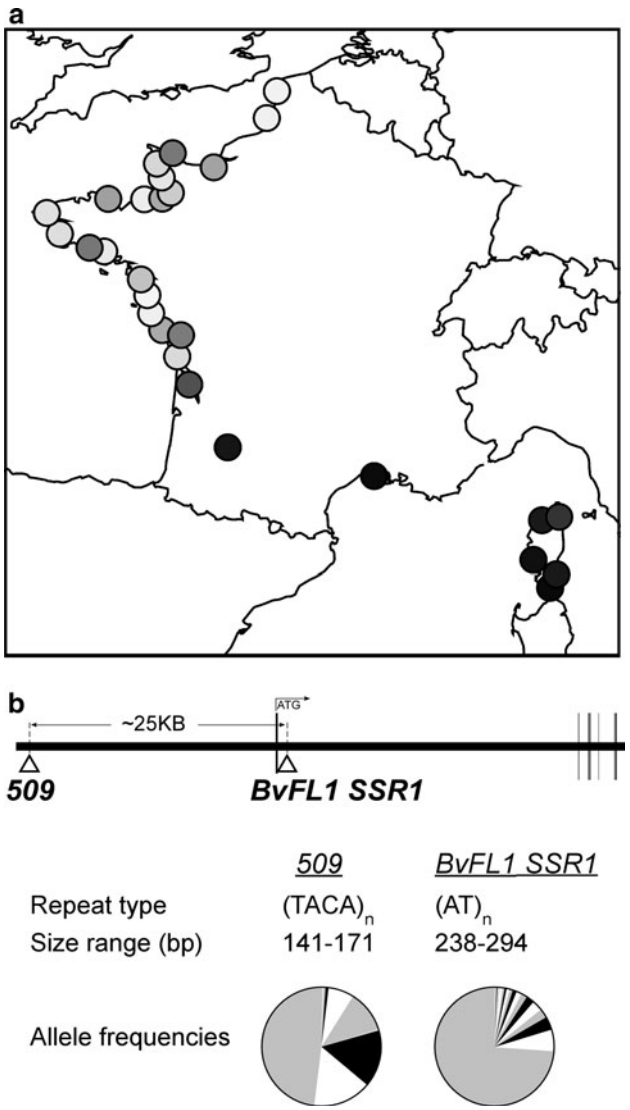

Fig. 1 **a** Geographic sampling of *Beta maritima* ssp. *maritima* accessions along coastal France. Site markers are *shaded* according to membership coefficient in one of two genetic clusters (Richards et al., in preparation). **b** Genomic region containing SSR loci linked to *BvFL1*, a locus of agronomic importance in sugar beet. *Triangles* mark the position of each SSR locus; *vertical lines* represent *BvFL1* exons. Translation initiation point is indicated above exon 1. Allele frequency spectra for neutral SSR locus *509* and selected locus *BvFL1 SSR1* shown using *pie charts*

BvFL1 SSR1 is located within the *BvFL1* gene, in intron 1, ~1 Kbp downstream from the start codon (Fig. 1b). Primer sequences for amplification were *509* (509.83F = TGCTCT CATCATCTTCTCCAATAG, 509.61R_{700} = ATATTT TTAGTGAATTTAGAAAG); *BvFL1 SSR1* (BvFL1 + 948F = ATGAAGTTCTAACCTTTATCACAA, BvFL1 + 1208R_{800} = AATGGTACGTGTATTATGAAACAT). PCR reactions contained 3 mM $MgCl_2$, 0.2 mM each dNTP, 0.05 units Taq DNA polymerase per µl, 2.5 pmol each primer, and 4 ng of DNA template. Cycling conditions consisted of an initial denaturation step at 95° for 5 min, followed by 30 cycles between 94°, 53°, and 72°, holding at each temperature for 45 s, then a final incubation at 72° for 10 min. Markers were visualized using a LI–COR 4200 DNA sequencer and scored using Saga GT software (LI–COR Biosciences, Lincoln, NE).

For all genotyped loci, PyPop (Lancaster et al. 2007) was used to perform the Ewens–Watterson–Slatkin (EWS) homozygosity test of neutrality (Slatkin 1994, 1996). This test evaluates the deviation of observed levels of homozygosity from levels expected under Hardy–Weinberg equilibrium. Higher than expected levels suggest directional or stabilizing selection (where a single allele is favored), and lower than expected levels suggest balancing selection (heterozygote advantage). A two-tailed test was used to distinguish the two alternative forms of selection from the null hypothesis of neutrality. For $\alpha = 0.05$, $p > 0.975$ indicated significant directional

selection, and $p < 0.025$ indicated significant balancing selection.

Distribution of neutral and selected variation at the agronomic locus BvFL1

We wished to determine which of the 28 *Beta vulgaris* ssp. *maritima* accessions were required in subsets that captured 100% of the allelic diversity at the neutral or selected target locus in *BvFL1*. Because many different subsets, each comprised of different accessions, might be "ideal" (i.e. contain all possible alleles), it is necessary to calculate a probability of occurrence for each accession. A Monte Carlo algorithm was used to determine constituents and calculate the probability of occurrence of each accession in idealized subsets. The algorithm proceeded as follows (with each target locus treated separately):

1. Assign all accessions to the subset.
2. Drop one accession, *A*, at random.
3. Determine the number of distinct alleles in the subset at the target locus. If all possible alleles are present, eliminate *A*. If some alleles have been lost, return *A* to the subset.
4. Repeat steps 2 and 3 until no further accessions can be eliminated without diminishing the number of alleles.
5. Record the identity of accessions included in the resulting, ideal subset.
6. Repeat steps 1–5 a total of 10,000 times. Calculate the probability of occurrence of each accession as the frequency it was recovered across replicates in the ideal subsets from step 5.

This Monte Carlo procedure was repeated, but with only 85% of total allelic diversity mandated to occur in the subset at step 3. Last, the procedure was applied to individuals, rather than accessions, to determine the probability of each individual occurring in a subset required to retrieve 85 or 100% of allelic diversity present at the target loci.

Core collection assembly

The M procedure (Schoen and Brown 1993), using the same search heuristic employed in MSTRAT (Gouesnard et al. 2001), was used to assemble subsets by choosing accessions such that the number of alleles recovered at reference loci was maximized. One thousand subsets were constructed for sizes ranging from 2 to 28 accessions. The number of alleles recovered at target loci was tabulated for each recommended subset. We modified the optimality criterion used in M (a simple count of distinct alleles at the reference loci) by normalizing allelic diversity to the total number of alleles observed at a locus across the data set (divide actual allele count by total possible alleles) and then standardizing the sum of the normalized allelic diversity values to the total number of reference loci examined (divide by number of loci). The MSTRAT search heuristic was then used to assemble subsets as before with one additional difference. MSTRAT does not support proper coding of codominant data; it treats each column of genotypic data as a distinct variable. The modified M procedure described, which we designate "M+", allows multiple columns to be assigned to a single locus and considered jointly during estimation of allelic diversity, permitting proper treatment of codominant data. Retention of allelic diversity at target loci under M+ was tabulated as for M so that relative performance could be compared.

Simulation of neutral and selected loci

Allelic variation at neutral and selected target loci was simulated under a coalescent model of evolution (Kingman 1982; Hudson 1983) using the software MSMS (Version 1.2.1; Ewing and Hermisson 2010). We assumed 50 segregating populations consisting of 1,000 individuals each, from which 10 diploid genotypes were sampled. The mutation parameter, $\theta = 4N_e\mu$, was varied randomly from 0 to 0.5, so that for a theoretical mutation rate (μ) of 1×10^{-8}, the effective population size (N_e) can be thought of as varying from 0 to 6.25×10^6, a plausible range for plant species. Loci included 1,000 segregating sites (conceptually, a 1,000-bp fragment) and the recombination parameter, ρ, was established such that the ratio ρ/θ varied randomly from 0 to 10, consistent with estimates of the recombination rate in several crop species (Morrell et al. 2006; Chen et al. 2008). A low uniform migration rate ($M = 4N_em$) of 0.01 was established by default between all populations to permit coalescence. To simulate more elaborate population structures, the default migration rate was modified between randomly chosen pairs of populations. The total number of such modifications varied from 0 to 9,900 (all possible unidirectional migration vectors for 50 populations), and the modified migration rate was drawn from a gamma distribution with a mean from 0 to 4 and a shape parameter from 0 to 10. This approach resulted in a wide diversity of possible population structures, from highly interconnected to highly isolated to complex combinations of the two.

Three categories of data sets were simulated: neutral reference loci, other neutral loci, and selected loci. Ten thousand distinct models were constructed and 50 loci were simulated for each model for each of the three categories of data. For a given model, all three data set categories had identical neutral parameters (i.e. identical migration matrices, θ and ρ). However, for the selected loci, three additional parameters were specified. The software MSMS uses a discrete, forward simulation approach for loci under

selection; hence the population size must be defined. For selected loci, the discrete population size was set to 50,000, which can be thought of as 50 populations of 1,000 individuals each, from which 500 individuals (10 per population) were eventually sampled. The strength of selection was varied at random from 1 to 1,000 (from "weak" to "strong" according to the MSMS manual) and codominance was assumed such that the selection strength on a heterozygote was half that of the homozygote at the selected site. We assumed that the selected site was fixed 0 to 10,000 generations in the past. For each locus, variable sites output by MSMS were interpreted as haplotypes which could be simplified into allelic states using the Perl utility ConStruct (Huelsenbeck and Andolfatto 2007). Because MSMS only returns variable sites, and our selected site was, by definition, fixed prior to the time of sampling, loci exhibited the signature of selection via the effect of linked polymorphism. Therefore, resulting allele frequency spectra at selected loci were dependent upon the interplay between randomly assigned values for strength of selection and ρ. EWS tests were performed to confirm that allele frequencies at neutral and selected loci conformed to expectations.

To consider whether findings from the single empirical data set suggest a general phenomenon, performance of the M+ optimization procedure was examined using the simulated data. Two optimized subsets were assembled for subset sizes ranging from 2 to 50 using the 50 neutral reference loci as a guide. A single locus was chosen from the set of 50 simulated loci, for each data set category. Allele counts were made at this single target locus for each of the two optimized subsets. The average allelic diversity retained was tabulated. This procedure was repeated, with all 50 simulated loci targeted. A comparison of the relative capability to recover variation at the neutral loci used to inform subset assembly, at other neutral loci evolving under the same model, and at selected loci, could then be made.

Results

Empirical data

In the *Beta vulgaris* ssp *maritima* data, at the 10 reference SSR loci, the number of distinct alleles ranged from three to 30, and allelic richness from 2.028 to 6.235, across the set of 28 accessions (Table 2). Target loci showed similar levels of diversity and richness. Locus *509*, ~24 Kbp upstream of the *BvFL1* coding region, had 8 alleles and an allelic richness of 3.721 while *BvFL1 SSR1*, located within intron 1, had 19 alleles and an allelic richness of 2.852. Target locus *509* was inferred to be neutral using the EWS test, while *BvFL1 SSR1*, although tightly linked to *509*, showed a statistically significant signature of directional selection. *BvFL1 SSR1* contained a single allele at high frequency, but many additional alleles at low frequency, whereas the allele count at *509* was lower and allele frequencies more uniform, as expected for selected and neutral loci (Fig. 1b). Only one of ten reference loci was inferred to be under selection using the EWS test; the remainder of loci were consistent with neutral expectations.

A Monte Carlo procedure was used to calculate the probability of occurrence of each accession in ideal subsets targeting either *509* or *BvFL1 SSR1*. The sets of accessions found in ideal subsets for neutral locus *509* were largely non-overlapping with accessions found in ideal subsets for *BvFL1 SSR1* (Fig. 2). The probability of occurrence of each accession in ideal subsets for the two distinct target genes was not correlated ($p = 0.629$). Relatively few

Table 2 Allelic diversity and evidence for selection at reference and target loci in *Beta vulgaris* ssp. *maritima*

	# Alleles	Allelic richness	EWS p value	Selection
Reference loci				
SB13[a]	9	3.166	0.419	Neutral
GTT1[b]	6	3.034	0.245	Neutral
FDSB1002[c]	17	5.122	0.151	Neutral
FDSB1005[d]	7	3.346	0.132	Neutral
FDSB1001[d]	17	3.084	0.986*	Directional
FDSB1026[d]	30	6.235	0.103	Neutral
FDSB1027[c]	17	3.986	0.861	Neutral
GCC1[b]	3	2.028	0.214	Neutral
GAA1[b]	6	2.080	0.769	Neutral
SB09[a]	7	2.839	0.375	Neutral
Target loci				
509	8	3.721	0.2003	Neutral
BvFL1 SSR1	19	2.852	0.9966**	Directional

* $p < 0.05$; ** $p < 0.01$

[a] Locus described in Richards et al. (2004)

[b] Locus described in Viard et al. (2002)

[c] Locus described in McGrath et al. (2007)

[d] Locus developed by V. Laurent, personal communication

Fig. 2 *Opposed bar graphs* indicate the frequency of occurrence of accessions in subsets required to contain all alleles at target loci. *Shaded circles* between graphs refer to sampling site shown in Fig. 1. Sites in the south appear at the *top*, northerly sites at the *bottom*. The accessions necessary to maximize diversity in a core collection differed depending on whether a neutral locus (*509*) or a selected locus (*BvFL1 SSR1*) was targeted

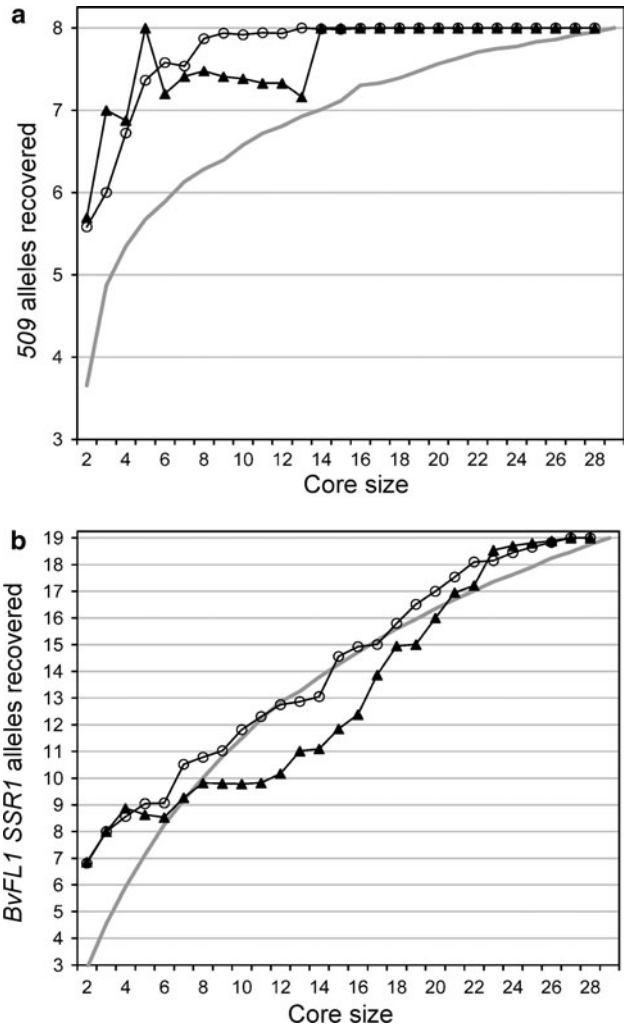

Fig. 3 Effect of core subset optimization algorithms on recovery of allelic diversity at neutral and selected target loci in *Beta vulgaris* ssp. *maritima*. Number of alleles retained in the core using random selection indicated with smooth *gray line*. Allelic retention using standard MSTRAT algorithm (M) shown with *solid triangles*. Improved algorithm (M+) shown with *open circles*. Both optimization algorithms performed well when targeting (**a**) neutral locus *509*, but failed when targeting (**b**) selected locus *BvFL1 SSR1*. M+ outperformed M in both cases

accessions (3–4) were necessary to retain all alleles at *509*. A single population from southern mainland France and one from Corsica were sampled most frequently. Retention of all alleles at *BvFL1 SSR1* required substantially more accessions (9–11), and they originated in sampling sites across the latitudinal range. Of the six accessions that contained private alleles at *BvFL1 SSR1* (and thus were always found in ideal subsets) two were never found in subsets targeting *509*, and the other four were only found infrequently. Likewise, a Corsican accession containing a private allele at *509* was never required in ideal subsets targeting *BvFL1 SSR1*. Similar patterns were observed when only 85% of alleles were required, although accessions containing private alleles then occurred with probabilities less than 1. When individuals rather than accessions were used the result was similar: *509* diversity was recovered most efficiently by sampling individuals from southern mainland France and Corsica; *BvFL1 SSR1* diversity was distributed in individuals across latitudes.

When applied to empirical data from *Beta vulgaris* ssp. *maritima*, core subsets assembled using either the M or M+ procedure contained more alleles than random subsets for the neutral locus *509* (Fig. 3a). In contrast, when selected locus *BvFL1 SSR1* was targeted, the same or fewer alleles than random were recovered (Fig. 3b). Recovery varied little across iterations (CV $<$ 15%). For locus *509*, on average across subset sizes, M and M+ procedures recovered 10 and 11% more alleles, respectively, than a random assembly strategy (up to 36% more alleles recovered for certain subset sizes). For *BvFL1 SSR1*, M recovered 3% fewer alleles than random. Although for some small subsets more alleles were found (up to 50% more),

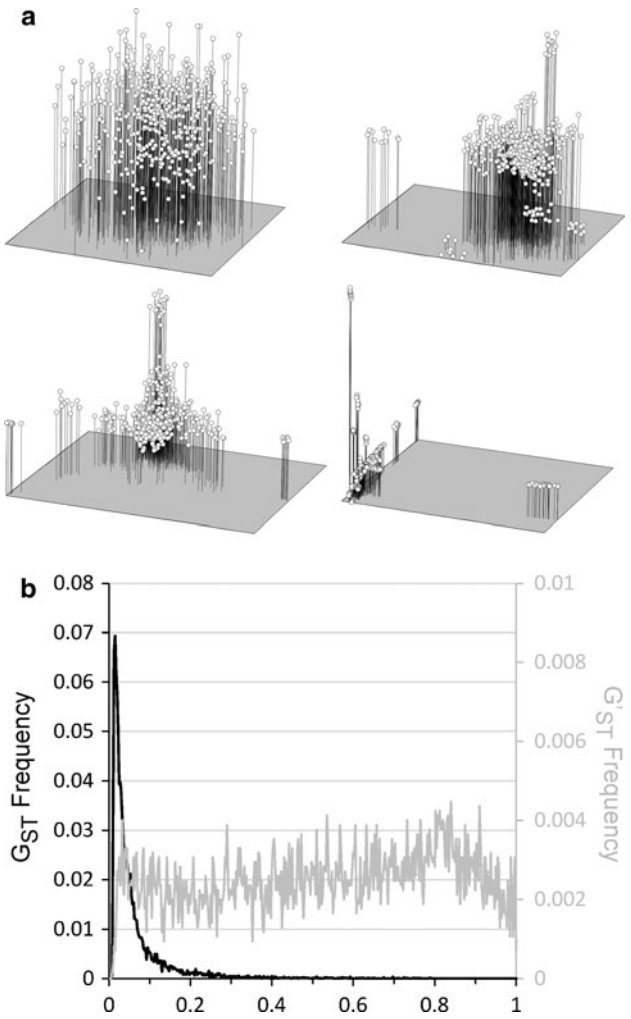

Fig. 4 **a** Four representative population structures from coalescent simulation of neutral reference loci, visualized using principal coordinate analysis. Models produced nearly panmictic to highly subdivided assemblages of populations. **b** Distribution of population differentiation observed in simulated data sets. Standardized metric G'_{ST} plotted in *gray*

other subsets contained as much as 24% fewer alleles. M+ recovered 4% more alleles than random; however, the difference was not significant ($p = 0.196$, paired t test). M performance was significantly worse than random ($p = 0.028$). M+ performed equal to or better than M for the majority of core sizes, regardless of target locus.

Simulated data

Ten thousand unique coalescent models were prepared. Random choice of migration parameters resulted in complete or nearly complete isolation for some populations in a minority of models. Forward simulations of selection in a coalescent context are memory intensive; thus models with highly restricted gene flow between some populations exceeded program memory limitations before all 50 loci could be simulated. Such incomplete models (15% of total) were discarded as were the corresponding models in the neutral simulations. In spite of removal of some models, the remaining 8,492 appeared to provide a diverse, random, sample from a relevant parameter space, as judged by the distribution of G'_{ST} values and ordination analyses of population structure (Fig. 4). Mean G_{ST} (± 1 SD) for neutral reference loci, other neutral loci, and selected target loci was 0.061 ± 0.081, 0.061 ± 0.080, and 0.063 ± 0.081, respectively (G'_{ST}: 0.528 ± 0.282, 0.527 ± 0.282, 0.069 ± 0.089). The range of G_{ST} values encountered across data sets was similar between data set categories (0–0.788). G'_{ST} values ranged from 0 to 1 for neutral loci, and from 0 to 0.759 for selected loci.

On average, 137.9 ± 94.2 (± 1 SD) segregating alleles were sampled from simulated neutral loci. Approximately 0.6% of loci were fixed. The maximum number of alleles sampled was 427. Selected target loci contained 15.9 ± 12.3 alleles. Five percent were fixed. The maximum number of segregating alleles observed at a selected locus was 170. Using the EWS test, 4.8% of variable, neutral loci had allele frequency distributions consistent with a history of directional selection, 2.7% with balancing selection, and the rest were neutral. The percentage of simulated neutral loci exhibiting a significant signature of selection was therefore just slightly higher than what might be expected given $\alpha = 0.05$. Ninety-five percent of loci simulated under a model with selection showed a significant signature of directional selection. Only one of the 404,565 variable simulated selected loci showed the signature of balancing selection. Hence, the simulation strategy produced data consistent with expectations (Type I error = 0.075, Type II error = 0.05).

Using M+, diversity was best retained when reference loci, those used for core optimization, were targeted. On average, for core sizes between 18 and 22, four additional alleles were retained relative to that retained when subsets were assembled at random (Fig. 5). Variation at neutral target loci that had not been used for optimization was not retained nearly as well. Fewer than two additional alleles were recovered relative to random, regardless of core size. Variation at selected loci was poorly retained. On average, less than one additional allele was recovered. The maximum degree of enrichment occurred at core size = 26, and was only 0.13 additional alleles.

The retention of target alleles in optimized cores was not correlated with values used for model parameters, although the variance in retention between models increased with increasing θ. The degree of population subdivision, on the other hand, as measured by G'_{ST}, was significantly correlated with allelic enrichment and with the total number of alleles. When allelic enrichment values were normalized by total alleles, a significant positive linear relationship remained between G'_{ST} and allelic enrichment for neutral

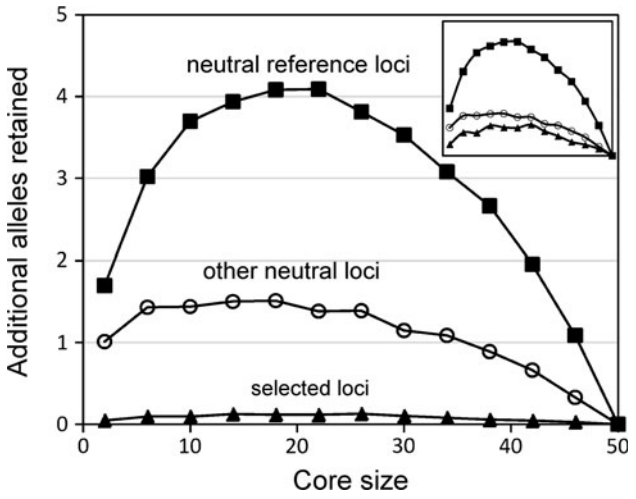

Fig. 5 Additional alleles retained in core subsets relative to randomly assembled subsets using M+ algorithm. On average, less than one additional allele was retained at simulated selected loci when using neutral loci to guide core assembly. *Inset* shows relative performance normalized to account for differences in the total number of alleles between categories of target loci

reference loci and other neutral loci, but not for selected loci ($p = 5.1 \times 10^{-60}$, 4.9×10^{-64}, 0.066; $R^2 = 0.276$, 0.145, 0.003). Therefore, as population subdivision increased, improved retention of neutral, but not selected diversity was found (Supplementary Figure 1). This relationship can be explained at a more fundamental level by examining the correlation in allele count between reference and target loci within populations. When 50 target loci were considered simultaneously, a strong, positive, curvilinear relationship was observed between neutral reference loci and neutral target loci (third order polynomial $R^2 = 0.598$) (Fig. 6a), whereas for target loci under selection a slight, positive, linear relationship was found ($R^2 = 0.180$, slope = 0.3) (Fig. 6b). For neutral targets, no models produced a negative correlation when G'_{ST} was greater than ~ 0.5. Negative correlations existed for selected targets even for models with $G'_{ST} \approx 1$. When only a single target locus was considered, the interlocus correlation between neutral reference and neutral target diversity remained positive (Fig. 6c), but the correlation

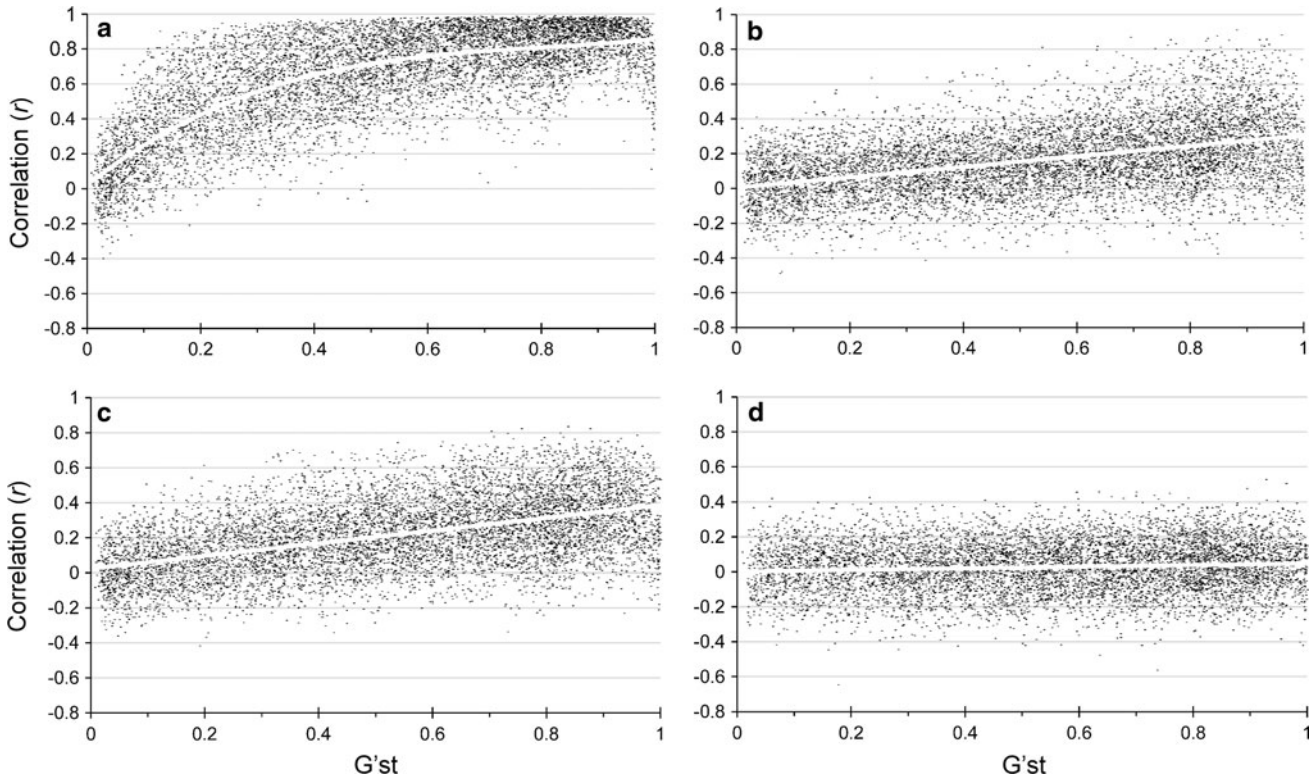

Fig. 6 Relationship between population subdivision and correlation in allelic diversity between loci. G'_{ST} based on neutral reference loci. Correlation coefficient (r) calculated between reference loci and (**a**) 50 neutral target loci, (**b**) 50 target loci under selection, (**c**) 1 neutral target locus, (**d**) 1 selected target locus. Ten thousand distinct coalescent models were sampled. *White line* indicates linear regression (third order polynomial regression for (**a**)). Increasing subdivision resulted in tighter interlocus correlation in allelic diversity, except when loci under selection were considered individually. Little opportunity exists to enhance variation at any individual selected locus using neutral loci to guide core collection assembly

between neutral reference and selected target diversity was near zero ($R^2 = 0.007$, slope = 0.046) (Fig. 6d).

Discussion

Efficient screening of diverse germplasm, especially wild germplasm (Tanksley and McCouch 1997; Gur and Zamir 2004), will be necessary to retrieve valuable phenotypes to combat emerging agricultural challenges (Kumar et al. 2010). In principle, utilization of core collections should accelerate the extraction of beneficial adaptations from genebanks by making the exploration of large germplasm collections for novel alleles more efficient. Inexpensive genotyping has made marker-based core collection optimization popular. There is, however, a theoretical roadblock to successful use of marker-based core collections for allele mining: the primary targets of allele mining projects—loci of agronomic importance—often exhibit patterns of allelic diversity consistent with a history of artificial or natural selection, while the types of loci used to inform core subset assembly are, more often than not, neutral loci.

Neutral loci are useful for describing the genetic structure of populations that has emerged as a consequence of historical demography and patterns of gene flow within a species. Patterns of polymorphism at selected loci may conflict with patterns at neutral loci because the movement of adaptive alleles among populations is not a passive process. A useful trait or adaptation can move rapidly through a set of populations (Slatkin 1976; Morjan and Rieseberg 2004). The number of neutral alleles segregating in a population, on the other hand, is primarily a consequence of its size (Kimura and Crow 1964), with a frequency spectrum consistent with principles of random sampling (Ewens 1972). The number and frequency spectrum of alleles at loci under selection depends upon the strength of selection at the selected site as well as recombination rates in the adjacent portion of the genome (Braverman et al. 1995; Fay and Wu 2000). The strength of selection for a particular trait value may vary widely among populations across a species' geographical range, whereas the effect of population size on variation at neutral and selected loci is fixed. Strength of selection and population size need not be correlated. Therefore, the distribution of variation at loci under selection will not necessarily be correlated with that at neutral loci. The distribution of allelic diversity at a locus under selection might only be predictable if the environmental factors acting to modify allele frequencies at the locus could be explicitly quantified across the sample, or set of accessions. Empirical data support these theoretical arguments. Using data from 29 species, McKay and Latta (2002) argued that a correlation between neutral markers and adaptive (selected) diversity is not expected. Their explanation for the lack of expected correlation is that (a) alleles at neutral loci are free to migrate, selected alleles are not, and (b) differentiation due to drift is slower than differentiation due to selection—neutral and selected loci will seldom be sampled in equilibrium.

Neutral and selected loci exhibit conflicting distributions of allelic diversity

We examined variation at two SSR loci linked to *BvFL1*, a vernalization-responsive gene in sugar beet which may have a role in controlling flowering time, an economically important trait in the crop (Reeves et al. 2007). Locus *BvFL1 SSR1*, located within the transcription unit (in intron 1), showed a strong signature of directional selection, while locus *509*, in spite of its proximity (24 Kbp upstream of the *BvFL1* start codon), exhibited a neutral pattern of genetic variation. Thus, two distinct evolutionary processes have affected the pattern of polymorphism at the two tightly linked loci. *BvFL1 SSR1* contained a single allele at high frequency, but many additional alleles at low frequency, whereas the allele count at *509* was lower and allele frequencies were more uniform, as expected for selected and neutral loci, respectively (Fig. 1b).

Using a Monte Carlo procedure, the frequency of occurrence of 28 *Beta vulgaris* ssp. *maritima* accessions in ideal core subsets for the two loci was computed. The accessions sampled most in ideal cores for locus *509* were not the same as for *BvFL1 SSR1* (Fig. 2). Thus, the distribution of allelic diversity across the sampled accessions differed markedly between neutral *509* and selected *BvFL1 SSR1*. While we present only a single empirical case, the finding is in agreement with theory and suggests that it may not be possible to simultaneously maximize retention of allelic diversity at both neutral and selected loci within the same core subset. This may hold true even if the neutral reference loci are tightly linked, both physically and statistically, to the agronomic locus of interest, as is the case here.

The M+ optimization procedure outperforms M

In order to come to a general understanding of the effect of core collection assembly using neutral reference loci on the retention of alleles at selected target loci, many data sets must be examined. The need for a core assembly procedure that is uniformly applicable across widely varying data sets, regardless of differences in the number of reference loci, or levels of variation at those loci, motivated our reexamination of the optimality criterion used in the most popular method for core assembly based on genotypic data, Schoen and Brown's (1993) M procedure.

In this study we introduce the M+ core assembly procedure. This modification of the M procedure is meant to address technical problems present in MSTRAT (Gouesnard et al. 2001) and PowerCore (Kim et al. 2007), as well as some conceptual issues. First, when performing M optimization, MSTRAT and PowerCore overestimate the total number of alleles in a data set by treating all loci as haploid. This weakens optimization, particularly for subsets made of individuals, because heterozygotes cannot be recognized and leveraged as such. Second, MSTRAT treats properly coded missing data as a distinct allele. These problems are simple to remedy by revising computer code (as has been done for M+), and do not reflect any problems with the M procedure per se.

A more fundamental problem with M is that it does not treat all loci equally. M uses a simple allele count as its optimality criterion. Loci with many alleles exert a greater influence on subset assembly than loci with few alleles by virtue of their greater potential to push the optimality criterion higher. M+ solves this by employing a normalization routine wherein the count of alleles at a locus in a subset is divided by the maximum possible number of alleles that would be observed at that locus if the entire data set were to be examined. This renders every locus equal in terms of its potential to elevate the optimality criterion. This improvement results in better estimation of genomewide patterns of polymorphism, because unusual, highly polymorphic portions of the genome are not emphasized during optimization. Second, because M uses a simple allele count, improvement in allele retention is not directly comparable between data sets. M+ solves this by dividing the normalized metric described above by the total number of reference loci in the data set. Thus, the diversity metric and optimality criterion used in M+ varies from 0 to 1, with equal contribution from all reference loci.

Using M+, recovery of target diversity was improved relative to M (Fig. 3). On average, across all subset sizes, M+ outperformed M for both neutral and selected target loci (although neither performed better than random for selected loci). For neutral locus *509*, M+ offered predictable performance across the range of subset sizes. The performance of M, on the other hand, was erratic, surpassing M+ for very small subsets ($n = 2, 3, 5$), but lagging for medium-sized subsets ($n = 6$–10). This erratic behavior is likely attributable to the non-standardized allele counts used for the optimality criterion in M.

Coalescent simulations of neutral and selected loci

Empirical data sets containing genotypes at both neutral reference loci and target loci of agronomic importance are currently rare. Coalescent simulations were performed to explore a large variety of different such data sets in an effort to produce a more general conclusion. Simulations can never fully model the complexities of natural population genetic processes; however, examination of the data sets simulated here suggests that they provide reasonable coverage of the range of variation in magnitude and apportionment of genetic diversity that might be found in the wild, or within germplasm collections. Because of the random approach used to assign parameter values, a broad diversity of population structures could be considered, encompassing scenarios from panmictic to highly structured (Fig. 4a). While simulated data cannot be expected to permit precise predictions for any given individual data set that might be the subject of an empirical study, these simulated data sets seem suitable for exploring average effects.

The mean level of variation observed within simulated subpopulations relative to total variation ($F_{ST} \approx G_{ST} = 0.06 \pm 0.08$) was, on average, lower than that observed in nature. Hamrick and Godt (1997) calculated $G_{ST} \approx 0.3$ and 0.2, for crops and wild species, respectively, using allozyme data. DNA-based RAPD data was similar (Nybom and Bartish 2000). Nevertheless, the range of values encountered in the simulated data sets (Fig. 4b) overlapped substantially with that observed across plants ($G_{ST} = 0.036$–0.510; 40% of models overlap). It is important to realize, however, that G_{ST} is a biased estimator of population differentiation, one whose value depends on the level of within-population homozygosity (Hedrick 2005). G_{ST} and related measures are not useful for comparing levels of differentiation between data sets (Jost 2008). When we use a standardized metric, G'_{ST}, we find little bias in our set of models toward any particular level of differentiation. Levels of population differentiation between $G'_{ST} = 0$ and 1 (mean $G'_{ST} = 0.53 \pm 0.28$) were uniformly explored (Fig. 4b).

Using the M+ core assembly algorithm, substantial improvements in the retention of diversity were only achieved for neutral target loci that were drawn from the set of reference loci (Fig. 5). Improvement in this category of loci is not likely to be of interest for users of germplasm collections because such loci do not impact crop improvement, nor are they indicative of adaptation to local environments (McKay and Latta 2002). Retention of variation at other neutral loci, those not used for optimization, was not so readily achieved, despite being simulated using the same model as the reference loci. Worse still was our ability to recover variation at selected loci. Although significantly better than random assembly, recovery of allelic diversity at selected loci was not meaningfully better because, on average, less than one additional allele was retained. Only for 14% of models was the average improvement in allelic recovery across core sizes greater than one additional allele. Although selected loci had fewer

alleles overall than neutral loci, poor retention was not a numerical artifact. When allelic retention was normalized by the maximum number of alleles possible for a given core size, the performance of selected target loci, and neutral target loci not used for optimization, became similar, and was still substantially worse than single targeted reference loci (Fig. 5, inset). On average, the (normalized) improvement seen for non-reference neutral target loci, and selected target loci, was 27 and 19%, respectively, of that observed when reference loci were used as target. Only occasionally would such average gains translate into the retention of additional alleles at any given target locus.

Marker-based core collections may be worse than random

Our results, both empirical and based on simulations, present a perspective that is counter to the prevailing one on the utility of core collections. Our results suggest that variation at a given locus of agronomic interest may not be elevated in core collections built using neutral loci, and may, in fact, be lower than collections assembled at random. The use of neutral marker–based core collections for allele mining may not expedite the discovery of useful variation at loci of agronomic importance. We arrive at this conclusion uncomfortably, because we have heretofore been generally supportive of the concept.

While at first glimpse this conclusion appears to contradict the critical studies on the topic, a more detailed reading reveals some points of agreement. Schoen and Brown (1993) examined the M strategy for its ability to enhance diversity at neutral target loci, noting that variation potentially useful in the future may be selectively neutral at present. They concluded that M might be especially useful for crops and inbreeding species, where the correlation between reference and target loci is greater due to elevated levels of genomewide linkage disequilibrium. Relatively lackluster performance (although still better than random) of M+ with non-reference, neutral target loci in our simulation might be attributed to a model that explicitly included migration and recombination, properties common in outcrossing species.

Schoen and Brown (1993) considered only nine empirical data sets, of which seven showed significantly improved allelic retention using M. Assuming an equal probability of retaining either more or less alleles than random, the binomial probability of observing seven or more positive results out of nine observations is 0.09, not statistically significant using the conventional $\alpha = 0.05$. Allelic retention was meaningfully improved (>1 additional allele) in five of the nine data sets, an equivocal result (binomial probability = 0.5). Schoen and Brown (1993) also calculated a metric of correlation of the levels of allelic diversity among loci (R). Six of eight data sets had a positive R value, while two (which failed to enrich using M) were zero or negative indicating no, or a negative, correlation in levels of diversity among loci. There is no significant difference between the mean of their empirical distribution of R values and a similarly sized distribution with mean zero and the same standard deviation (t test, $p = 0.216$). Hence, the specific data sets examined by Schoen and Brown (1993) do not, in a statistical sense, support the idea that allelic diversity is correlated among neutral loci. However, we do not dispute this notion, both on theoretical grounds, and based upon the results of our simulation. Our results provide statistical rigor to Schoen and Brown's hypothesis, and demonstrate, in addition, that meaningful enrichment might also be obtained at neutral target loci in outcrossing species.

Our results appear to conflict more directly with Bataillon et al. (1996) who, using a different simulation strategy, found that (a) retention of selected alleles was improved using neutral marker loci and the M strategy, and (b) retention of neutral alleles was similar to alleles at selected loci regardless of the subset assembly procedure used. We found substantial differences in retention of neutral compared with selected target loci and that selected target locus diversity could not be meaningfully elevated (Fig. 5). Moreover, they predicted that M, in particular, would never perform more poorly than random assembly, arguing that "For this to happen, there would have to be a negative correlation between diversity at marker and selected loci" (p. 415). Monte Carlo simulations undertaken here demonstrate precisely this kind of negative correlation with real data (Fig. 2), and, as a consequence, M performed worse than random for the selected target locus *BvFL1 SSR1* (Fig. 3b). With simulated data, we observed a negative correlation in allele count between neutral reference and selected target loci for 21% of the models examined. In contrast, for neutral target loci, only 2.6% of models showed a negative correlation (Fig. 7a, 50 target loci considered simultaneously). Likewise, the mean correlation of reference loci diversity with selected target loci was much lower ($r = 0.16 \pm 0.20$) than with non-reference, neutral target loci ($r = 0.65 \pm 0.27$).

Based upon simulated data, only a very slight positive correlation between neutral and selected variation is predicted. Furthermore, that slight positive correlation is an average effect, calculated across many loci. If target loci are considered individually (as they would be in an allele mining experiment) the ability to predict, based on neutral reference loci, which populations will retain elevated variation at the target locus worsens. A negative correlation between diversity at neutral reference loci and diversity at a single target locus under selection was observed for 42% of models (Fig. 7b). The average correlation was very

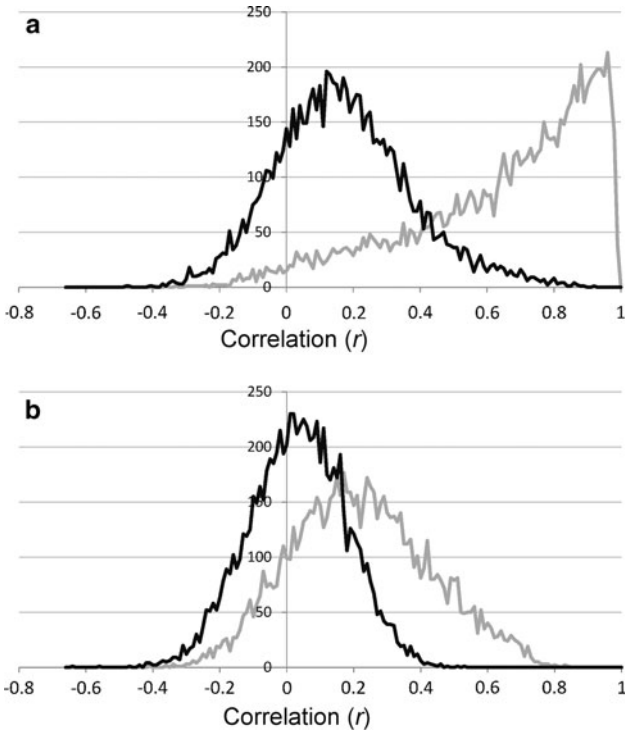

Fig. 7 Correlation in allelic diversity between simulated neutral and selected loci. *X* axis is correlation coefficient *r*; *Y* axis is frequency of occurrence. Distribution of *r* values for neutral target loci shown in *gray*, target loci under selection in *black*. **a** 50 target loci. **b** A single target locus. Allelic diversity is positively correlated for neutral loci—populations exhibiting high allelic diversity at neutral reference loci were also highly diverse at neutral target loci. The correlation in allelic diversity between neutral reference loci and selected target loci was weak, and frequently negative

close to zero (0.03 ± 0.14). Therefore, one should not expect the retention of diversity at an individual selected target locus to be improved by using neutral loci to guide core collection assembly. Random assembly would be predicted to achieve a similar level of enrichment. When individual neutral loci were targeted the mean correlation was higher (0.22 ± 0.21) and only 16% of models showed a negative correlation. Thus, consistent with Schoen and Brown (1993), our simulation predicts that it should be straightforward to enhance variation in a set of unsampled neutral loci in core collections. A single neutral target locus will be more difficult, and an individual locus with a history of selection nearly impossible using conventional approaches.

Because this conclusion is strongly supported by the simulated data, yet runs completely counter to the broadly-accepted findings presented in Bataillon et al. (1996), we shall make a detailed effort to explain the apparent discrepancy. We propose that the conflict between our study and that of Bataillon et al. (1996) is most likely attributable to differences in how selected loci were simulated and measured. Their model considered three selective environments, among which 18 demes (populations) were distributed, forming a structured set of populations. Migration occurred under a finite island model. The simulation strategy was based on that of David et al. (1993) and used selection on individual fitness as a way, in principle, to produce a data set containing selected loci. This was accomplished using fitness curves, which described the fitness contribution of all alleles segregating at each of the ten physically unlinked loci. Individual fitness was calculated additively, as the sum of fitness values for all alleles present in an individual's multilocus genotype. The deviation of this value from the population mean was then used to choose parents for the subsequent generation. Under these conditions, M performed significantly better than random for selfers and when migration among demes was prohibited for outcrossers.

It is unclear to us the extent to which this protocol would produce independent loci exhibiting the characteristic allele frequency spectra that form the statistical signature of selection (EWS test values for selected loci were not presented). Selection performed thusly in finite, genetically isolated populations will not necessarily drive individual loci to their respective optima because of the establishment of strong linkage disequilibrium between loci (Fisher 1930; Kimura 1956; Lewontin 1964). When fitness is determined in an additive manner across two or more loci, selection generates negative linkage disequilibrium (where favored alleles at one locus become associated with unfavored alleles at another locus), which reduces the response to selection, and decreases the rate of change of allele frequencies—the so–called 'Hill–Robertson effect' (Felsenstein 1965; Hill and Robertson 1966). Beneficial alleles caught in inferior combinations disappear due to selection (and/or drift) before they can be recombined into superior multilocus genotypes (Fisher 1930; Muller 1932). Under these conditions, gene flow between isolated populations (or more precisely, recombination) is critical for maintaining the additive genetic variation necessary to achieve the optimal allelic combination (Felsenstein 1974; David et al. 1993; Martin et al. 2006).

Bataillon et al. (1996) only found a correlation in diversity between neutral and selected loci for selfing populations and for genetically isolated outcrossing populations. These are precisely the cases where selection is expected to be inefficient in moving allele frequency distributions away from neutral expectations. Among the categories of mating system and demography they considered, theory predicts that loci in outcrossing populations with migration would exhibit the strongest statistical signature of selection. No correlation between neutral and selected diversity was found for this category of populations. In other words, a correlation was observed when loci under selection likely did not show the signature of

selection and was not observed when the loci likely did show the signature of selection.

We suspect that excessive population subdivision may be the source of the strong correlation between neutral and putative selected diversity observed by Bataillon et al. (1996). They considered a single population structure and eight demographic conditions. Using a two-population coalescent model, Hudson and Coyne (2002) showed that after 4–7 N_e generations without gene flow, 50% of loci will exhibit reciprocal monophyly, where, for a given locus, all alleles found within one population will be more closely related to one another than they are to any alleles found at the same locus in a second population (95% of loci reciprocally monophyletic at 9–12 N_e generations). This is a sufficient level of differentiation to be considered distinct species under a variety of species definitions (Donoghue 1985; Nixon and Wheeler 1990; Baum and Shaw 1995; Mallet 1995). With simulated population sizes ranging from 100 to 3,000 (N_e expected to be lower, especially for selfing populations) and 5,000 generations without gene flow, levels of differentiation between some populations in their simulation surely reached magnitudes expected for distinct species. Hudson and Coyne's (2002) work predicts that for $N_e = 500$, 95% of loci would be reciprocally monophyletic in 4,500–6,000 generations. The data sets considered by Bataillon et al. (1996) must therefore have been extremely subdivided and may not appropriately represent the levels of population differentiation expected within a typical species-level germplasm collection. Elevated genomewide linkage disequilibrium is the hallmark of strong population subdivision (Nei and Li 1973; Ohta 1982; practical applications: Pritchard et al. 2000; England et al. 2010) and would seem to be a likely source of the correlation in level of diversity they observed.

Last, when calculating enrichment, Bataillon et al. (1996) measured allelic retention additively, at all ten selected target loci simultaneously. Simultaneous consideration of ten target loci is not necessarily pertinent to allele mining, which is generally concerned with a single target locus. Such a procedure may, however, be appropriate for understanding the retention of diversity at the QTL underlying polygenic traits of agronomic importance.

Therefore, we believe that the simulation model used in Bataillon et al. (1996) is not applicable to the issue of allele mining for agronomically important genetic variation. It likely created a situation where the signature of selection could not be achieved due to extreme population subdivision and the Hill–Robertson effect. We suggest that the correlation in allelic diversity observed between neutral loci and those purported to be under selection was, in fact, the strong correlation in diversity that is expected to arise between all loci in genetically isolated populations. Nevertheless, our results do not entirely contradict their conclusions. We agree with their finding that increased subdivision results in better performance for M (or M+) (Fig. 6; Supplementary Figure 1). The two studies, however, pertain to different units of conservation and imply different conservation objectives. Using the coalescent rather than a discrete, time-forward simulation, we were able to explore a vastly larger set of possible population structures, containing levels of differentiation consistent with what might be observed within single species, the unit of conservation in most gene banking efforts.

The findings presented here are not optimistic with respect to the utility of neutral marker-based core collections for allele mining. They should not, however, be surprising. Natural selection is a powerful force that can drive allele frequencies in patterns counter to neutral expectations. Very little of the historical signature written across the genome onto neutral loci need remain on individual selected genes in the same genome. Elimination of obvious redundancy and basic stratification of collections using geography or habitat type prior to subset assembly should still generate modest efficiency increases. Likewise, core sets based on phenotypic surveys of the agronomic traits of interest might be enhanced for allelic diversity at the genes underlying those traits (although this remains to be demonstrated). Additionally, it may be possible to rely on an assembly procedure like M+ to improve the retention of diversity, on average, across a targeted gene network, or at a group of loci underlying a polygenic trait, wherein each locus may have experienced different selection pressures during its history. A deeper understanding of the effect of historical patterns of selection, genetic drift, and population subdivision on genomewide linkage disequilibrium will be necessary (1) to determine whether correlations in diversity between neutral and selected loci are expected for a given germplasm collection and (2) to leverage that correlation for the formation of core subsets optimized for allele mining.

Acknowledgments We thank Ann Fenwick for genotyping the reference loci, and Gayle Volk and Dale Lockwood for comments on the manuscript.

References

Backes G, Graner A, Foroughi–Wehr B, Fischbeck G, Wenzel G, Jahoor A (1995) Localization of quantitative trait loci (QTL) for agronomic important characters by the use of a RFLP map in barley (*Hordeum vulgare* L.). Theor Appl Genet 90:294–302

Balfourier F, Roussel V, Strelchenko P, Exbrayat-Vinson F, Sourdille P, Boutet G, Koenig J, Ravel C, Mitrofanova O, Beckert M, Charmet G (2007) A worldwide bread wheat core collection arrayed in a 384–well plate. Theor Appl Genet 114:1265–1275

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

Baum DA, Shaw KL (1995) Genealogical perspectives on the species problem. In: Hoch PC, Stephenson AG (eds) Experimental and molecular approaches to plant biosystematics. Missouri Botanical Garden, St. Louis, pp 289–303

Bernacchi D, Beck–Bunn T, Eshed Y, Lopez J, Petiard V, Uhlig J, Zamir D, Tanksley S (1998) Advanced backcross QTL analysis in tomato. I. Identification of QTLs for traits of agronomic importance from Lycopersicon hirsutum. Theor Appl Genet 97:381–397

Bhullar NK, Street K, Mackay M, Yahlaoul N, Keller B (2009) Unlocking wheat genetic resources for the molecular identification of previously undescribed functional alleles at the *Pm3* resistance locus. Proc Natl Acad Sci USA 106:9519–9524

Bhullar NK, Zhang Z, Wicker T, Keller B (2010) Wheat gene bank accessions as a source of new alleles of the powdery mildew resistance gene Pm3: a large scale allele mining project. BMC Plant Biol 10:88

Bishop GJ, Harrison K, Jones JDG (1996) The tomato *Dwarf* gene isolated by heterologous transposon tagging encodes the first member of a new cytochrome P450 family. Plant Cell 8:959–969

Börner A (2006) Preservation of plant genetic resources in the biotechnology era. Biotechnol J 1:1393–1404

Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140:783–796

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

Caicedo AL (2008) Geographic diversity cline of *R* gene homologs in wild populations of *Solanum pimpinellifolium*. Am J Bot 95:393–398

Casa AM, Mitchell SE, Hamblin MT, Sun H, Bowers JE, Paterson AH, Aquadro CF, Kresovich S (2005) Diversity and selection in sorghum: simultaneous analyses using simple sequence repeats. Theor Appl Genet 111:23–30

Chapman MA, Pashley CH, Wenzler J, Hvala J, Tang S, Knapp SJ, Burke JM (2008) A genomic scan for selection reveals candidates for genes involved in the evolution of cultivated sunflower (*Helianthus annuus*). Plant Cell 20:2931–2945

Charlesworth B, Charlesworth D, Barton NH (2003) The effects of genetic and geographic structure on neutral variation. Annu Rev Ecol Evol S 34:99–125

Chen H, Morrell PL, de la Cruz M, Clegg MT (2008) Nucleotide diversity and linkage disequilibrium in wild avocado (*Persea americana* Mill.). J Hered 99:382–389

Crossa J, Burgueño J, Dreisigacker S, Vargas M, Herrera–Foessel SA, Lillemo M, Singh RP, Trethowan R, Warburton M, Franco J, Reynolds M, Crouch JH, Ortiz R (2007) Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure. Genetics 177:1889–1913

David JL, Savy Y, Brabant P (1993) Outcrossing and selfing evolution in populations under directional selection. Heredity 71:642–651

Donoghue MJ (1985) A critique of the biological species concept and recommendations for a phylogenetic alternative. Bryologist 88:172–181

El Mousadik A, Petit RJ (1996) High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* (L.) Skeels] endemic to Morocco. Theor Appl Genet 92:832–839

England PR, Luikart G, Waples RS (2010) Early detection of population fragmentation using linkage disequilibrium estimation of effective population size. Conserv Genet 11:2425–2430

Escribano P, Viruel MA, Hormaza JI (2008) Comparison of different methods to construct a core germplasm collection in woody perennial species with simple sequence repeat markers: a case study in cherimoya (Annona cherimola, Annonaceae), an underutilised subtropical fruit tree species. Ann Appl Biol 153:25–32

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Popul Biol 3:87–112

Ewing G, Hermisson J (2010) MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics 26:2064–2065

Fay JC, Wu C-I (2000) Hitchhiking under positive Darwinian selection. Genetics 155:1405–1413

Felsenstein J (1965) The effect of linkage on directional selection. Genetics 52:349–363

Felsenstein J (1974) The evolutionary advantage of recombination. Genetics 78:737–756

Fisher RA (1930) The genetical theory of natural selection. Clarendon Press, Oxford

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

Frankel OH (1984) Genetic perspectives of germplasm conservation. In: Arber W, Illmensee K, Peacock WJ, Starlinger P (eds) Genetic manipulation: impact on man and society. Cambridge University Press, Cambridge, pp 161–170

Frankham R (1996) Relationship of genetic variation to population size in wildlife. Conserv Biol 10:1500–1508

González-Martínez SC, Wheeler NC, Ersoz E, Nelson CD, Neale DB (2007) Association genetics in *Pinus taeda* L. I. wood property traits. Genetics 175:399–409

Goudet, J (2001) FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3). Software distributed by the author

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

Gur A, Zamir D (2004) Unused natural variation can lift yield barriers in plant breeding. PLoS Biol 2:e245

Hamrick JL, Godt MJW (1997) Allozyme diversity in cultivated crops. Crop Sci 37:26–30

Hedrick PW (2005) A standardized genetic differentiation measure. Evolution 59:1633–1638

Hill WG, Robertson A (1966) The effect of linkage on limits to artificial selection. Genet Res 8:269–294

Hoisington D, Khairallah M, Reeves T, Ribaut J-M, Skovmand B, Taba S, Warburton M (1999) Plant genetic resources: what can they contribute toward increased crop productivity? Proc Natl Acad Sci USA 96:5937–5943

Hudson RR (1983) Properties of a neutral allele model with intragenic recombination. Theor Popul Biol 23:183–201

Hudson RR, Coyne JA (2002) Mathematical consequences of the genealogical species concept. Evolution 56:1557–1565

Huelsenbeck JP, Andolfatto P (2007) Inference of population structure under a Dirichlet process model. Genetics 175:1787–1802

Hyten DL, Song Q, Zhu Y, Choi I–Y, Nelson RL, Costa JM, Specht JE, Shoemaker RC, Cregan PB (2006) Impacts of genetic bottlenecks on soybean genome diversity. Proc Natl Acad Sci USA 103:16666–16671

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

Johal GS, Briggs SP (1992) Reductase activity encoded by the *HM1* disease resistance gene in maize. Science 258:985–987

Johal GS, Balint–Kurti P, Weil CF (2008) Mining and harnessing natural variation: a little MAGIC. Crop Sci 48:2066–2073

Jost L (2008) G_{ST} and its relatives do not measure differentiation. Mol Ecol 17:4015–4026

Kaur N, Street K, Mackay M, Yahiaoui N, Keller B (2008) Molecular approaches for characterization and use of natural disease resistance in wheat. Eur J Plant Pathol 121:387–397

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

Kimura M (1956) A model of a genetic system which leads to closer linkage by natural selection. Evolution 10:278–287

Kimura M, Crow JF (1964) The number of alleles that can be maintained in a finite population. Genetics 49:725–738

Kingman JFC (1982) On the genealogy of large populations. J Appl Prob 19A:27–43

Kumar GR, Sakthivel K, Sundaram RM, Neeraja CN, Balachandran SM, Rani NS, Viraktamath BC, Madhav MS (2010) Allele mining in crops: prospects and potentials. Biotechnol Adv 28:451–461

Lancaster AK, Single RM, Solberg OD, Nelson MP, Thomson G (2007) PyPop update—a software pipeline for large-scale multilocus population genomics. Tissue Antigens 69(s1):192–197

Latha R, Rubia L, Bennett J, Swaminathan MS (2004) Allele mining for stress tolerance genes in *Oryza* species and related germplasm. Mol Biotechnol 27:101–108

Le Cunff L, Fournier–Level A, Laucou V, Vezzulli S, Lacombe T, Adam–Blondon A-F, Boursiquot J–M, This P (2008) Construction of nested genetic core collections to optimize the exploitation of natural diversity in *Vitis vinifera* L. subsp. sativa. BMC Plant Biol 8:31

Lewontin RC (1964) The interaction of selection and linkage. II. Optimum models. Genetics 50:757–782

Mallet J (1995) A species definition for the modern synthesis. Trends Ecol Evol 10:294–299

Martin G, Otto SP, Lenormand T (2006) Selection for recombination in structured populations. Genetics 172:593–609

McGrath JM, Trebbi D, Fenwick A, Panella L, Schulz B, Laurent V, Barnes, Murray SC (2007) An open–source first–generation molecular genetic map from a sugarbeet × table beet cross and its extension to physical mapping. The Plant Genome 47:S27–S44

McKay JK, Latta RG (2002) Adaptive population divergence: markers, QTL and traits. Trends Ecol Evol 17:285–291

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

Morjan CL, Rieseberg LH (2004) How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. Mol Ecol 13:1341–1356

Morrell PL, Toleno DM, Lundy KE, Clegg MT (2006) Estimating the contribution of mutation, recombination and gene conversion in the generation of haplotypic diversity. Genetics 173:1705–1723

Muller HJ (1932) Some genetic aspects of sex. Am Nat 66:118–138

Nei M, Li W-H (1973) Linkage disequilibrium in subdivided populations. Genetics 75:213–219

Nielsen R (2005) Molecular signatures of natural selection. Annu Rev Genet 39:197–218

Nixon KC, Wheeler QD (1990) An amplification of the phylogenetic species concept. Cladistics 6:211–223

Nybom H, Bartish IV (2000) Effects of life history traits and sampling strategies on genetic diversity estimates obtained with RAPD markers in plants. Perspect Plant Ecol Evol Syst 3:93–114

Ohta T (1982) Linkage disequilibrium due to random genetic drift in finite subdivided populations. Proc Natl Acad Sci USA 79:1940–1944

Prada D (2009) Molecular population genetics and agronomic alleles in seed banks: searching for a needle in a haystack? J Exp Bot 60:2541–2552

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

Reed DH, Frankham R (2001) How closely correlated are molecular and quantitative measures of genetic variation? A meta–analysis. Evolution 55:1095–1103

Reeves PA, He Y, Schmitz RJ, Amasino RM, Panella LW, Richards CM (2007) Evolutionary conservation of the *FLC*–mediated vernalization response: evidence from the sugar beet (*Beta vulgaris*). Genetics 176:295–307

Richards CM (2004) Molecular technologies for managing and using gene bank collections. In: MC de Vicente (ed) The evolving role of genebanks in the fast developing field of molecular genetics. Issues Genet Resour 11:13–18. IPGRI, Rome, Italy

Richards CM, Brownson M, Mitchell SE, Kresovich S, Panella L (2004) Polymorphic microsatellite markers for inferring diversity in wild and domesticated sugar beet (*Beta vulgaris*). Mol Ecol Notes 4:243–345

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

Schoen DJ, Brown AHD (2001) The conservation of wild plant species in seed banks. Bioscience 51:960–966

Slatkin M (1976) The rate of spread of an advantageous allele in a subdivided population. In: Karlin S, Nevo E (eds) Population genetics and ecology. Academic Press, New York, pp 767–780

Slatkin M (1994) An exact test for neutrality based on the Ewens sampling distribution. Genet Res 64:71–74

Slatkin M (1996) A correction to the exact test based on the Ewens sampling distribution. Genet Res 68:259–260

Tanksley SD, McCouch SR (1997) Seed banks and molecular maps: unlocking genetic potential from the wild. Science 277:418–423

Tenaillon MI, U'Ren J, Tenaillon O, Gaut BS (2004) Selection versus demography: a multilocus investigation of the domestication process in maize. Mol Biol Evol 21:1214–1225

Viard F, Bernard J, Desplanque B (2002) Crop–weed interactions in the *Beta vulgaris* complex at a local scale: allelic diversity and gene flow within sugar beet fields. Theor Appl Genet 104:688–697

Vigouroux Y, McMullen M, Hittinger CT, Houchins K, Schulz L, Kresovich S, Matsuoka Y, Doebley J (2002) Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. Proc Natl Acad Sci USA 99:9650–9655

Walsh B (2008) Using molecular markers for detecting domestication, improvement, and adaptation genes. Euphytica 161:1–17

Walters C, Volk GA, Richards CM (2008) Genebanks in the post-genomic age: emerging roles and anticipated uses. Biodiversity 9:68–71

Wang M, Allefs S, van den Berg RG, Vleeshouwers VGAA, van der Vossen EAG, Vosman B (2008) Allele mining in *Solanum*: conserved homologues of *Rpi–blb1* are identified in *Solanum stoloniferum*. Theor Appl Genet 116:933–943

Whitham S, Dinesh–Kumar SP, Choi D, Hehl R, Corr C, Baker B (1994) The product of the tobacco mosaic virus resistance gene *N*: similarity to toll and the interleukin–1 receptor. Cell 78:1101–1115

Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159

Wright SI, Gaut BS (2005) Molecular population genetics and the search for adaptive evolution in plants. Mol Biol Evol 22:506–519

Xiao J, Li J, Yuan L, Tanksley SD (1996) Identification of QTLs affecting traits of agronomic importance in a recombinant inbred

population derived from a subspecific rice cross. Theor Appl Genet 92:230–244

Yamasaki M, Tenaillon MI, Vroh Bi I, Schroeder SG, Sanchez–Villeda H, Doebley JF, Gaut BS, McMullen MD (2005) A large–scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. Plant Cell 17:2859–2872

Zhu Q, Zheng X, Luo J, Gaut BS, Ge S (2007) Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: severe bottleneck during domestication of rice. Mol Biol Evol 24:875–888