Theor Appl Genet (2011) 123:693–704
DOI 10.1007/s00122-011-1619-3

ORIGINAL PAPER

# Genetic diversity and population structure in cultivated sunflower and a comparison to its wild progenitor, *Helianthus annuus* L

**J. R. Mandel · J. M. Dechaine · L. F. Marek · J. M. Burke**

Received: 1 November 2010 / Accepted: 14 May 2011 / Published online: 3 June 2011


**Abstract** Crop germplasm collections are valuable resources for ongoing plant breeding efforts. To fully utilize such collections, however, researchers need detailed information about the amount and distribution of genetic diversity present within collections. Here, we report the results of a population genetic analysis of the primary gene pool of sunflower (*Helianthus annuus* L.) based on a broad sampling of 433 cultivated accessions from North America and Europe, as well as a range-wide collection of 24 wild sunflower populations. Gene diversity across the cultivars was 0.47, as compared with 0.70 in the wilds, indicating that cultivated sunflower harbors roughly two-thirds of the total genetic diversity present in wild sunflower. Population structure analyses revealed that wild sunflower can be subdivided into four genetically distinct population clusters throughout its North American range, whereas the cultivated sunflower gene pool could be split into two main clusters separating restorer lines from the balance of the gene pool. Use of a maximum likelihood method to estimate the contribution of the wild gene pool to the cultivated sunflower germplasm revealed that the bulk of the cultivar diversity is derived from two wild sunflower population genetic clusters that are primarily composed of individuals from the east-central United States, the same general region in which sunflower domestication is believed to have occurred. We also identified a nested subset of accessions that capture as much of the allelic diversity present within the sampled cultivated sunflower germplasm collection as possible. At the high end, a core set of 288 captured nearly 90% of the alleles present in the full set of 433, whereas a core set of just 12 accessions was sufficient to capture nearly 50% of the total allelic diversity present within this sample of cultivated sunflower.

Communicated by A. Bervillé.

**Electronic supplementary material** The online version of this article (doi:10.1007/s00122-011-1619-3) contains supplementary material, which is available to authorized users.

J. R. Mandel · J. M. Burke (✉)
Miller Plant Sciences, University of Georgia, Athens, GA 30602, USA
e-mail: jmburke@uga.edu

J. M. Dechaine
University Way Central Washington University, 400 E., Ellensburg, WA 98926, USA

L. F. Marek
Iowa State University, G202 Agronomy, Ames, IA 50011, USA

## Introduction

The value of germplasm collections for ongoing crop improvement was first recognized by Nikolai Vavilov, who argued that "the practical plant-breeder uses this material as bricks with which he must construct new forms" (Vavilov 1940). Indeed, such resources have regularly served as a source of novel alleles for ongoing plant breeding efforts in a variety of species (Tanksley and Nelson 1996; Acquaah 2006). Unlocking the full potential of crop germplasm collections, however, requires an understanding of the amount and distribution of genetic variation contained within them. To this end, we analyzed the genetic diversity present within the primary gene pool of sunflower (*Helianthus annuus* L.) based on a broad sampling of cultivated accessions, i.e., cultivars, from North America and Europe, as well as a range-wide collection of wild sunflower populations.

Cultivated sunflower is a globally important oilseed crop and an important source of confectionery seeds and

ornamental flowers. Originally domesticated from the self-incompatible common sunflower (also *H. annuus*) approximately 4,000 years ago in what is now the central United States (Heiser et al. 1969; Smith 1989; Rieseberg and Seiler 1990; Crites 1993; Harter et al. 2004; Smith 2006), cultivated sunflower—which is self-compatible—was used by Native Americans as a source of edible seeds as well as for a variety non-food applications (e.g., as a source of dye for textiles and for ceremonial purposes); (Soleri and Cleveland 1993). In the early sixteenth century, sunflower was taken to Europe by Spanish explorers where it was first grown as an ornamental plant (Putt 1997). During the eighteenth century and beyond, sunflower was increasingly used as a source of vegetable oil, and breeding efforts focused primarily on improving oil yield. This was particularly true in the late nineteenth century through the mid-twentieth century in Eastern Europe, where sunflower was grown on a large scale (Pustovoit 1964).

Ultimately, the germplasm that formed the basis of the modern 'oilseed' sunflower gene pool was brought back to North America, and the first commercial high-oil sunflowers were grown in the United States in the 1960s (Putt 1997). These early high-oil cultivars were almost exclusively open-pollinated varieties, though attention turned to hybrid production—along with a concomitant focus on the production of inbred lines—in the 1970s, primarily due to the higher yields and greater disease resistance afforded by hybrids (Robertson and Burns 1975). During this time, confectionery-type cultivars were also being developed and grown in the United States (Robertson and Morrison 1977), but oilseed production soon surpassed that of non-oilseed sunflowers, and breeders increasingly shifted their focus toward developing improved oilseed cultivars (Cheres and Knapp 1998).

A key development during the transformation of sunflower into a hybrid oilseed crop, which now accounts for 80–85% of sunflower production in the United States, was the discovery of cytoplasmic male sterility (Leclercq 1969) along with fertility restoration genes (Kinman 1970). These discoveries resulted in the development of two distinct breeding types within the sunflower gene pool. Restorer (R) lines are homozygous for dominant nuclear restorer alleles at one or more fertility restorer (*Rf*) loci. In contrast, maintainer (B) lines are homozygous for the non-restorer allele(s) on a normal (i.e., male-fertile) cytoplasmic background (Fick and Miller 1997). Each B line is paired with a corresponding A line which has the same nuclear genotype, but carries a male-sterile cytoplasm. Crosses between B and A lines "maintain" the male-sterile A line due to the maternal inheritance of the cytoplasm. Fertile hybrids are thus produced when R lines are crossed with A lines, and the maintenance of the R lines and A/B lines as distinct breeding pools helps to maximize heterosis. Confectionery inbred lines have also been developed following this same general approach, albeit from an apparently narrow genetic base (Cheres and Knapp 1998).

It should be clear from above that cultivated sunflower has experienced a complex and varied evolutionary history. Unfortunately, the impact of these events on genetic diversity within the sunflower gene pool has not been fully explored. While researchers have previously sought to genetically characterize the sunflower gene pool, such studies have generally focused on a single agronomic type of sunflower (i.e., either oilseed or confectionery—e.g., Hongtrakul et al. 1997; Dong et al. 2007) and/or on a relatively limited number of accessions (e.g., Cronn et al. 1997; Fusari et al. 2008; Yue et al. 2009). In the present study, we analyzed genetic diversity and population structure in a broad sampling of oilseed and confectionery sunflower accessions from North America and Europe, as well as in a diverse set of wild *H. annuus* populations from across the species range in North America. All sampled individuals were genotyped using a genome-wide collection of simple sequence repeat (SSR) markers derived from expressed sequence tags (ESTs). The resulting data allowed us to investigate the amount and distribution of genetic diversity within the cultivated sunflower gene pool in the context of its wild progenitor and to identify a nested subset of cultivars that captures the maximal amount of overall allelic diversity.

## Materials and methods

### Plant materials and genotyping

Seed from 433 cultivated *H. annuus* accessions was obtained from the USDA North Central Regional Plant Introduction Station (NCRPIS) and the French National Institute for Agricultural Research (INRA). These 433 accessions included numerous inbred lines, a selection of Native American landraces, and historically important open-pollinated varieties including high-oil Eastern European cultivars, as well as some accessions from elsewhere in the world. Moreover, at least one accession was chosen from each of the 10 "core clusters" as defined by Brothers and Miller (1999). This previously described core collection included 112 accessions including landraces and open-pollinated varieties and was subdivided into "core clusters" based upon 20 morphological descriptors.

Eight seeds from each accession were germinated in petri dishes on moist filter paper and a 10% solution of BanRot (Scotts Company, Marysville, OH). Following germination, seedlings were planted in two-inch pots in the greenhouse, and DNA was extracted from three to four individuals per line using a CTAB extraction protocol

(Doyle and Doyle 1987). Each accession was then culled to a single individual, which was transferred to a 3-gallon pot, grown to maturity and self-pollinated to establish inbred lines for future studies. For comparison to the progenitor of cultivated sunflower, we selected a sample of 96 wild *H. annuus* individuals from 24 populations (4 individuals per population) from across the species range in the United States, Mexico, and Canada (Online Resource Table 1). A map representing the locations of the 24 wild, *H. annuus* populations is provided in Fig. 1. Populations were sampled from the central United States, representing the native range of wild sunflower (Heiser et al. 1969) and where it is presently most common, as well as locations in western Canada and northern Mexico. Seed from these wild populations was likewise obtained from NCRPIS, and DNA was extracted as above.

All cultivated accessions were assigned to one of twelve categories based on their origin (USDA or INRA), breeding history (maintainer [B] lines = HA; restorer [R] lines = RHA), and agronomic use (oil vs. non-oil). Because an oil or confectionery designation was not available for the INRA accessions, these were divided into INRA-derived B and R lines (denoted INRA-HA and INRA-RHA, respectively). For the USDA accessions, the following categories were defined: HA non-oil, HA oil, RHA non-oil, RHA oil, non-oil-introgressed, oil-introgressed, landrace, open-pollinated variety (OPV), other non-oil, and other oil (Online Resource Table 2; USDA 2006). Accessions designated 'non-oil' are either the confectionery type, or could not be clearly defined as being the oil type. The 'introgressed' categories included both oil and non-oil accessions with a recent history of wild *H. annuus* introgression, as indicated by the available pedigree information (e.g., Beard 1982; Korell et al. 1992). The landrace category was comprised of the following seven Native American landraces: Arikara, Havasupai, Hidatsa, Hopi, Mandan, Maíz Negro, and Seneca. The OPV category included named sunflower accessions that represent open-pollinated varieties of the pre-hybrid era of sunflower breeding, including Jupiter, Manchurian, Jumbo, VIR 847, Mammoth, etc. (BS Hulke, USDA-ARS, pers. comm.). The 'other oil' and 'other non-oil' categories included accessions of each type for which a B vs. R designation could not be made.

All DNA samples (three to four individuals per cultivated accession and four per wild population) were genotyped using 34 EST-SSRs chosen from the larger set of markers employed by Chapman et al. (2008). These 34 markers were selected based on presumptive neutrality (i.e., they showed no evidence of having been targeted by selection during sunflower domestication or improvement) and genetic map position. The end result was broad genomic coverage, with one marker on either arm of each of the 17 sunflower linkage groups (Online Resource

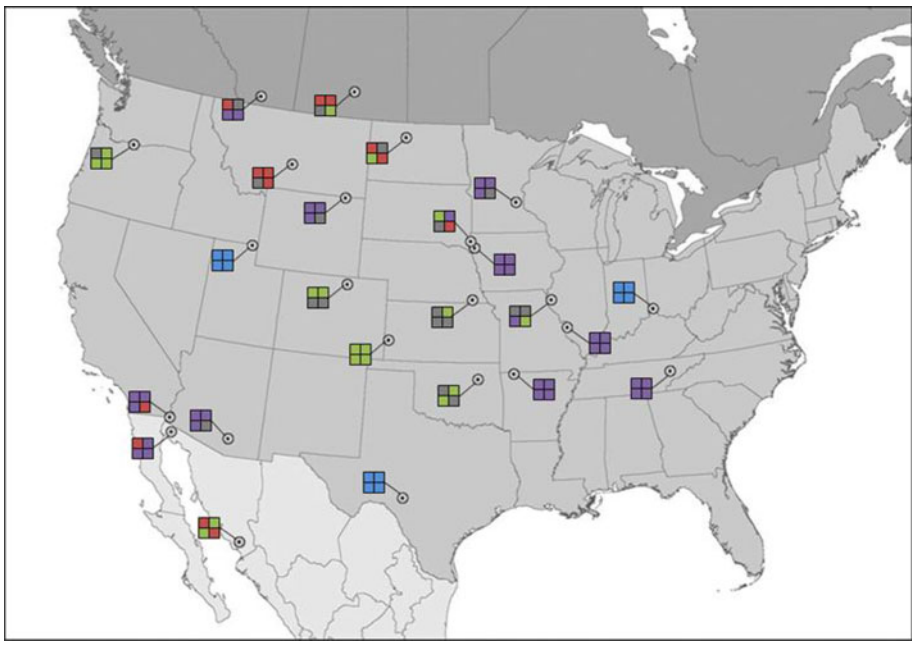

**Fig. 1** Map of the collection sites of the 24 wild *Helianthus annuus* populations (4 individuals per population) used in this study. These populations were selected to cover the native range of the species, which mainly encompasses the central United States. The *colored squares* alongside each population indicate membership of each of the four sampled individuals in one of the four population genetic clusters identified using STRUCTURE. Individuals with greater than 66% membership in a particular cluster were assigned to that cluster. If an individual did not meet this criterion, it was classified as admixed. *Blue*, cluster 1; *red*, cluster 2; *green*, cluster 3; *purple*, cluster 4; *gray*, admixed (see text for additional details)

Table 3). EST-SSR genotyping was performed using the fluorescent labeling protocol of Schuelke (2000) as modified by Wills et al. (2005). PCR was performed in a total volume of 20μL containing 5 ng of template DNA, 30 mM Tricine pH 8.4-KOH, 50 mM KCl, 2 mM $MgCl_2$, 125 μM of each dNTP, 0.2 μM M13 Forward (-29) sequencing primer labeled with either HEX, 6-FAM or NED, 0.2 μM reverse primer, 0.02 μM forward primer and 2 units of *Taq* polymerase. The PCR conditions were as follows: 3 min at 95°C; 10 cycles of 30 s at 94°C, 30 s at 65°C and 45 s at 72°C, annealing temperature decreasing to 55°C by 1°C per cycle, followed by 30 cycles of 30 s at 94°C, 30 s at 55°C, 45 s at 72°C, followed by 20 min at 72°C.

Amplicons were diluted 1:30 in deionized water and visualized using an ABI 3730xl DNA sequencer (Applied Biosystems, Foster City, CA) with MapMarker 1000 ROX size standards (BioVentures, Murfreesboro, TN) included in each lane to allow for accurate fragment size determination. Alleles were called using the software package GeneMarker v. 1.51 (SoftGenetics, State College, PA).

## Population genetic analyses and relatedness

Measures of genetic diversity, including mean number of alleles per locus, observed heterozygosity, and unbiased gene diversity (Nei 1978) were calculated across the total sample (three to four individuals per accession or population) of cultivated and wild sunflower using GenAlEx v. 6.1 (Peakall and Smouse 2006). All subsequent analyses of the cultivated accessions were based upon a single individual per cultivar (the one selected for selfing and inbred line establishment; see above) to reduce computational intensity and because observed heterozygosity was generally quite low (see "Results"). In restricting the sample size in the remaining analyses to one individual per accession, we recognize some genetic diversity occurring within accessions may be been overlooked. Rarefaction was used to compare allelic diversity of wild versus cultivated sunflower and to account for differences in sample size within these two groups (Hurlbert 1971; Petit et al. 1998; Kalinowski 2004). This analysis was performed using HP-Rare 1.0 (Kalinowski 2005) to estimate allelic richness based on the genotypic data for all 34 EST-SSRs from the 529 samples (i.e., the 96 wild + 433 cultivated individuals, which were analyzed separately). This same approach was also used to compare the allelic richness among the different cultivar categories (i.e., HA-Non-Oil, HA-Oil, RHA-Non-Oil, RHA-Oil, etc.). Two separate input files were created: one used for estimating allelic richness in the wilds and cultivars and one used for analysis of the cultivars alone. Program parameters were set according to the recommendations by Kalinowski, and the program was run with, and without, the assumption of missing data (2004; 2005).

Population structure in wild and cultivated sunflower was investigated using the Bayesian, model-based clustering algorithm implemented in the software package STRUCTURE (Pritchard et al. 2000a). Briefly, individuals were assigned to $K$ population genetic clusters based on their multi-locus genotypes. Clusters were assembled so as to minimize intra-cluster Hardy–Weinberg and linkage disequilibrium and, for each individual, the proportion of membership in each cluster was estimated. This analysis did not rely on prior population information (i.e., USEPOPINFO was turned off). The population structure present in our full dataset was first assessed by evaluating the wild and cultivated sunflower samples together in the STRUCTURE analysis ($N = 529$). The wild ($N = 96$) and cultivated ($N = 433$) subsets were then analyzed separately. For each analysis, $K = 1$–12 population genetic clusters were evaluated with 5 runs per $K$ value, and the probability values were averaged across runs for each cluster. For each run, the initial burn-in period was set to 50,000 with 100,000 MCMC iterations. Because a preliminary analysis using the admixture model and either correlated or independent allele frequencies revealed no appreciable differences between approaches, all analyses were performed assuming independent allele frequencies. The most likely number of clusters was then determined using the Delta$K$ method of Evanno et al. (2005).

Population structure was also investigated via analysis of molecular variation (AMOVA; Excoffier et al. 1992), as implemented in GenAlEx, which was used to hierarchically partition genetic variation and estimate $F_{ST}$ (Wright 1951). Initially, the extent of differentiation between the wilds and cultivars was estimated, followed by a pairwise analysis amongst the 12 cultivars classes. In all cases, statistical significance (i.e., $H_0$ = no genetic differentiation among the populations) was determined by performing 1,000 permutations. Note that an equivalent analysis amongst pairs of wild populations was not performed due to limited within-population sampling.

In order to investigate the genetic constitution of cultivated sunflower with respect to the wild sunflower gene pool, a mixture stock analysis was used to estimate the contribution of the wild gene pool to the cultivated germplasm surveyed herein. Mixture stock analyses estimate the proportion of the "mixed" stock, or group, which likely came from each of several given source populations. This approach used an unconditional maximum likelihood (UML) method (Pella and Milner 1987; Smouse et al. 1990) using the "mixstock" package (Bolker et al. 2003) in the "R" programming environment (R Development Core Team 2011). UML methods are preferable over conditional maximum likelihood methods because the latter assume that the true source frequencies are equal to the sampled frequencies, such that unsampled genotypes can cause

problems. In contrast, UML approaches allow for genotypic sampling error (Smouse et al. 1990; Bolker et al. 2003). In this analysis, the potential source populations were defined as the four wild sunflower population genetic clusters identified in the STRUCTURE analysis (see Results) and were limited to just those individuals with greater than two-thirds membership in a particular cluster (i.e., individuals with intermediate levels of membership in multiple clusters were excluded from consideration). A total of 13 mixed stock analyses were performed, where the "mixed stocks" were the total sample of cultivated germplasm as well as the 12 cultivar classes separately. In each case, 95% confidence intervals of the source contributions were obtained via 1,000 non-parametric bootstrap resamplings.

Genetic relationships amongst the cultivated sunflower accessions were also investigated graphically via principal coordinates (PCO) analysis using GenAlEx. A standard genetic distance matrix (Nei 1978) was constructed based on the multi-locus genotypes of the full sample of 433 individuals. This distance matrix was then used for the PCO analysis, and the first two principal coordinates were graphed in two-dimensional space. To further assess relationships amongst the cultivars included in this study, a neighbor-joining tree was constructed using the natural log transformation of the proportion of shared alleles distance (lnPSAD). Distances were calculated using PowerMarker (Liu and Muse 2005), and the phylogenetic tree was constructed using the neighbor-joining algorithm, NEIGHBOR, in the computer software package PHYLIP v. 3.68 (Felsenstein 2005). This algorithm was employed due to its fast computational speed, which was necessary due to the large number of individual data points. The program FigTree v. 1.3.1 (Morariu et al. 2008) was then used to draw the resulting tree.

### Construction of a core germplasm collection for sunflower

The simulated annealing algorithm of PowerMarker (Liu and Muse 2005) was used to identify a nested core set of cultivated sunflower accessions that captures the maximal allelic diversity based on the full set of 433 cultivated individuals. This algorithm maximizes diversity by choosing subsets of individuals that contain the greatest possible number of equally frequent alleles—i.e., the algorithm does not simply choose rare alleles to maximize allelic diversity. PowerMarker was first used to identify a core set of 12 individuals that captured the most allelic diversity out of the full set. Note that this analysis was based solely upon the genotypic data, and thus did not consider membership in the named categories defined above. This analysis was then repeated for a larger sets of individuals while constraining the algorithm to include the original 12 (e.g., the second subset comprised the 12 original individuals plus 36 more selected from the remaining 421 individuals). This process was repeated to produce nested sets of 12, 48, 96, 192, and 288 individuals, each representing unique cultivated accessions, that contained maximal allelic diversity. At each level, the selection procedure was repeated 100 times. Variation amongst replicates was minimal, with the majority of replicates comprising identical subsets. When differences were found, the replicate with the highest number of alleles was chosen. For each nested core set (i.e., 12, 48, 96, 192, and 288), the observed and expected heterozygosity was calculated, and the proportion of overall allelic diversity present in the subset of interest was estimated as the number of alleles present in that subset divided by the total number of alleles present in the entire sample of 433 individuals.

## Results

### Genetic diversity of the cultivated sunflower germplasm collection

In total, we surveyed 1,729 cultivated *H. annuus* plants corresponding to 433 cultivated accessions using 34 EST-SSRs. Gene diversity, or expected heterozygosity ($H_e$), across the total 1,729 sampled plants was $0.47 \pm 0.032$ (mean $\pm$ standard error), and the average observed heterozygosity ($H_o$) per plant was $0.076 \pm 0.0090$ (Table 1). The average number of alleles per locus was $6.8 \pm 0.37$, and the average intra-accession (3–4 plants per accession) observed heterozygosity was $0.059 \pm 0.0080$. Genetic diversity statistics were also calculated within each of the 12 named cultivar categories using one individual per accession. As expected, categories containing accessions with a history of wild introgression had the highest observed heterozygosity values. However, after adjusting for differences in sample sizes via rarefaction, allelic richness values did not differ among the various cultivar categories (Table 2; ANOVA, $df = 11$, $F = 0.64$, $P = 0.79$).

In addition to surveying genetic diversity across the publicly available cultivated sunflower germplasm, we investigated how well this diversity represents that which is present in wild sunflower populations from across the species range in the United States, Mexico, and Canada. Gene diversity in our sample of 96 wild individuals (4 individuals $\times$ 24 populations) was high, with mean $H_e = 0.70 \pm 0.031$ and the average observed heterozygosity was $H_o = 0.29 \pm 0.029$. On average, the genetic diversity of the cultivated germplasm was roughly 67% (0.47/0.70) of that observed in the wild. Across the 34 loci, the cultivars harbored fewer total alleles (230 vs. 328) and

**Table 1** Genetic diversity statistics for wild and cultivated sunflower

| Type | *N* | *A* | % Total *A* | $H_e$ | $H_o$ |
|---|---|---|---|---|---|
| Total cultivated | 1,729 | 267 | – | 0.47 | 0.076 |
| Total wilds | 96 | 328 | – | 0.70 | 0.286 |
| All 433 lines | 433 | 230 | 100 | 0.47 | 0.065 |
| Core 288 | 288 | 201 | 87.4 | 0.47 | 0.065 |
| Core 192 | 192 | 183 | 79.6 | 0.46 | 0.068 |
| Core 96 | 96 | 154 | 70.0 | 0.45 | 0.050 |
| Core 48 | 48 | 134 | 58.2 | 0.45 | 0.054 |
| Core 12 | 12 | 105 | 45.7 | 0.41 | 0.064 |

*N* number of plants sampled, *A* number of alleles observed, *% Total A* percent alleles of the 433 cultivated accessions, $H_e$ Nei's unbiased gene diversity averaged across loci, $H_o$ observed heterozygosity averaged across loci

**Table 2** Genetic diversity statistics for cultivated sunflower accessions grouped according to the 12 categories described in the "Materials and methods"

| Category | *N* | *A* | $A_g \pm$ SE | $H_e$ | $H_o$ |
|---|---|---|---|---|---|
| INRA-HA | 56 | 122 | 2.71 (0.14) | 0.43 | 0.037 |
| INRA-RHA | 30 | 105 | 2.77 (0.12) | 0.40 | 0.021 |
| HA-non-oil | 61 | 131 | 2.64 (0.10) | 0.46 | 0.053 |
| HA-oil | 78 | 134 | 2.47 (0.13) | 0.42 | 0.044 |
| RHA-non-oil | 27 | 102 | 2.61 (0.11) | 0.42 | 0.056 |
| RHA-oil | 83 | 130 | 2.72 (0.12) | 0.40 | 0.030 |
| Non-oil introgressed | 12 | 101 | 2.69 (0.11) | 0.42 | 0.217 |
| Oil introgressed | 23 | 117 | 2.75 (0.11) | 0.44 | 0.219 |
| Landrace | 7 | 83 | 2.71 (0.14) | 0.44 | 0.142 |
| OPV | 21 | 108 | 2.65 (0.13) | 0.42 | 0.132 |
| Other non-oil | 15 | 98 | 2.55 (0.11) | 0.42 | 0.069 |
| Other oil | 20 | 105 | 2.54 (0.13) | 0.44 | 0.087 |

*N* number of plants sampled, *A* number of alleles observed, $A_g$ allelic richness using the rarefaction method $\pm$ the standard error, $H_e$ Nei's unbiased gene diversity averaged across loci, $H_o$ observed heterozygosity averaged across loci

Data here was calculated on the single individual per cultivar, see "Materials and methods"

occupied a narrower range of allele sizes as compared with the wild sample (Online Resource Figure 1). Of the 230 alleles present in the cultivars, 87 were private, or unique to the cultivated germplasm. In contrast, wild sunflower had 186 private alleles out of the 328 total alleles. The cultivated collection and wild collection had 143 alleles in common. After adjusting for differences in sample sizes using the rarefaction method, allelic richness per locus in the wild was significantly higher than in the cultivars: $9.2 \pm 0.49$ versus $5.2 \pm 0.49$ (paired *t* test, $P < 0.0001$).

## Population structure and comparison of wild *Helianthus annuus* and cultivated sunflower

With regard to population structure in both wild and cultivated sunflower (i.e., 96 wilds + 433 cultivars, with a single individual representing each cultivated accession), the Delta*K* method of Evanno et al. (2005) provided support for the presence of two genetically distinct clusters (i.e., $K = 2$; Online Resource Figure 2), which corresponded to the wild versus cultivar distinction. In other words, all wild individuals showed ≥95% membership in one cluster, whereas all cultivars exhibited ≥95% membership in the other cluster (data not shown). The separate analysis of the 96 wild individuals provided support for the existence of $K = 4$ clusters (Figs. 1, 2; Online Resource Figure 2). The separate analysis of the 433 cultivated individuals provided greatest support for $K = 2$ clusters (Fig. 3; Online Resource Figure 2). These two genetically distinct subgroups largely corresponded to the restorer-oil (RHA-oil) lines versus the remainder of the gene pool, though the relationship was not exact.

In terms of population structure between recognized groups in this study, $F_{ST}$ as estimated from AMOVA was 0.22 ($P < 0.01$) between the wilds ($N = 96$) and cultivars ($N = 433$). Pairwise $F_{ST}$ amongst the 12 cultivar classes ranged from 0.016 to 0.183 with the highest values occurring between the INRA-RHA lines and the balance of the gene pool, as well as between the RHA-oil lines and the remainder of the cultivated gene pool (see Online Resource Table 5 for all pairwise $F_{ST}$ values along with *P* values).

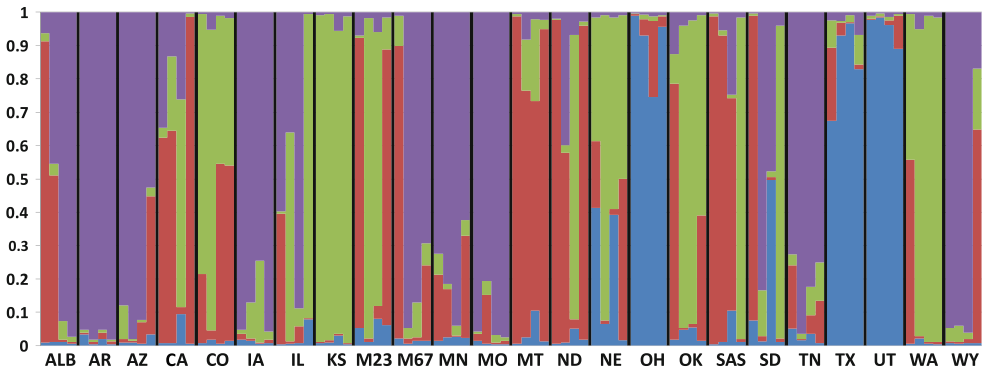


**Fig. 2** STRUCTURE plot of 96 wild *H. annuus* individuals with $K = 4$ clusters. *Blue*, cluster 1; *red*, cluster 2; *green*, cluster 3; *purple*, cluster 4. The *y*-axis shows the proportion membership into the various clusters. Each *colored vertical bar* represents a single wild individual and the 4 individuals from each of the 24 sampled populations are grouped together. *Vertical black bars* have been included as visual separators between the populations. The populations are arranged alphabetically, as in Online Resource Table 1, and the color codes correspond to those used in Figure 1. Online Resource Table 4 contains a complete listing of the geographic composition of each of the four clusters

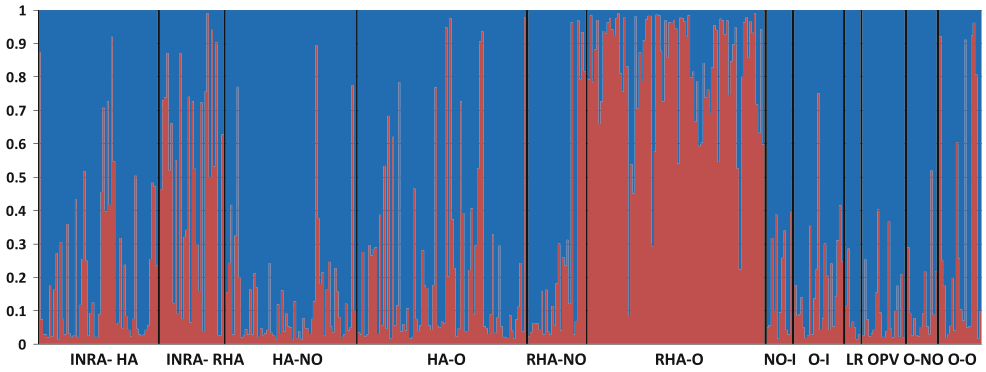


**Fig. 3** STRUCTURE plot of the 433 cultivated sunflower accessions with $K = 2$ clusters. *Blue*, cluster 1; *red*, cluster 2. The *y*-axis shows the proportion membership into the various clusters. This analysis was based on a single individual per accession using genotypic data from 34 EST-SSRs (see "Methods" for details). The *vertical black bars* have been included as visual separators between the named categories. *INRA-HA* INRA-derived HA lines, *INRA-RHA* INRA-derived RHA lines, *HA-NO* HA-non-oil lines, *HA-O* HA-oil lines, *RHA-NO* RHA-non-oil lines, *RHA-O* RHA-oil lines, *NO-I* non-oil introgressed, *O-I* oil introgressed, *LR* landraces, *OPV* open pollinated varieties, *NO* other non-oil lines, *O* other oil lines

The mixture stock analysis revealed that the majority of the source contributions for the cultivated gene pool (analyzed as one mixture as well as the 12 cultivar classes separately) were from wild clusters 3 and 4. Table 3 gives the UML point estimates for each of the four sources (wild clusters 1–4). The UML point estimates were consistently higher for cluster 3, however, bootstrap confidence intervals generally overlapped with cluster 4. Shared symbols in Table 3 indicate overlapping 95% confidence intervals as determined by 1,000 bootstraps. Figure 4 plots the point estimates from each wild cluster and confidence intervals for the total cultivated germplasm ($N = 433$). Online Resource Figure 3 also plots the point estimates for each source contribution along with the bootstrap confidence intervals.

Inspection of the PCO plot—which was simplified by refining our classification to include the following five categories: INRA-RHA, Landraces, RHA-oil, the Core12, and a category that contained all remaining cultivars—revealed extensive overlap amongst categories (Fig. 5). However, the INRA-RHA and RHA-oil lines were generally separated from the balance of the cultivated germplasm along coordinate two, with these lines typically having positive values. In contrast, the neighbor-joining tree showed little in the way of clear differentiation amongst accessions within the cultivated sunflower gene pool (Online Resource Figure 4).

## Nested core set

The simulated annealing algorithm of PowerMarker identified core subsets of 12, 48, 96, 192, and 288 individuals (Table 1). Online Resource Table 2 lists the accessions belonging to each of these subsets. The largest core set ($N = 288$) captured 87.4% (201/230 alleles) of the total

**Table 3** Estimated source (i.e., wild clusters 1–4) contributions to the total cultivated germplasm and the 12 cultivar classes based on the unconditional maximum likelihood (UML) analysis

| | Wild cluster 1 | Wild cluster 2 | Wild cluster 3 | Wild cluster 4 |
|---|---|---|---|---|
| Total | $0.124^{+}$ | $0.134^{+}$ | 0.427* | 0.315* |
| INRA-HA | $0.078^{+}$ | $0.167^{+}$ | 0.415* | 0.340* |
| INRA-RHA | $0.004^{+}$ | 0.233* | 0.417* | 0.346* |
| HA-non-oil | $0.171^{+}$ | $0.086^{+}$ | 0.438* | 0.305* |
| HA-oil | $0.183^{+}$ | $0.106^{+}$ | 0.413* | 0.298* |
| RHA-non-oil | $0.123^{\$}$ | $0.131^{+}$ | 0.475* | $0.271^{+,*}$ |
| RHA-oil | $0.060^{+}$ | $0.176^{+}$ | 0.463* | 0.301* |
| Non-oil introgressed | $0.245^{\$,*}$ | $0.138^{\$}$ | 0.356* | 0.261* |
| Oil introgressed | $0.176^{+}$ | $0.118^{+}$ | 0.416* | $0.290^{+,*}$ |
| Landrace | $0.164^{+}$ | $0.074^{+}$ | 0.441* | 0.321* |
| OPV | $0.160^{+}$ | $0.102^{+}$ | 0.366* | 0.372* |
| Other non-oil | $0.165^{+}$ | $0.074^{+}$ | 0.440* | 0.321* |
| Other oil | $0.145^{+}$ | $0.134^{+}$ | 0.395* | 0.326* |

Shared symbols indicate overlapping 95% confidence intervals as determined by 1,000 bootstrap replicates

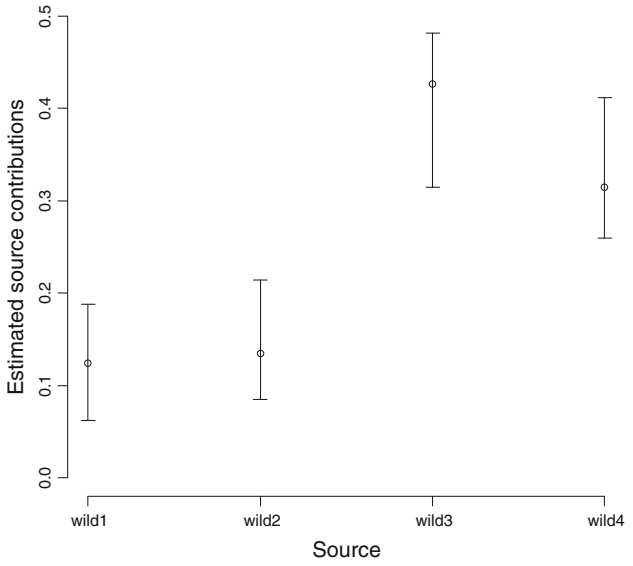


**Fig. 4** Unconditional maximum likelihood (UML) estimated source (wild clusters 1–4) contributions to the total cultivated germplasm. *Error bars* indicate 95% confidence intervals as determined by 1,000 bootstrap replicates

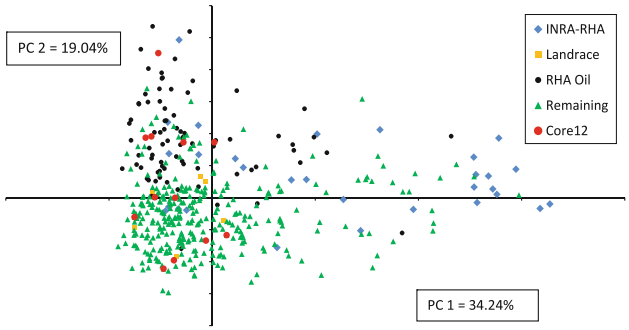


**Fig. 5** Principal coordinates (PCO) analysis of 433 cultivated sunflower accessions based on genotypic data from 34 EST-SSRs. This analysis was based on a single individual per accession (see "Methods" for details). In order to view the figure more easily, the named categories were condensed as follows: INRA, all INRA-derived lines; Landrace; RHA-oil; Core12, smallest core subset (see text for details); and Remaining, all remaining individuals, including HA-non-oil, HA-oil, RHA-non-oil, non-oil introgressed, oil introgressed, OPV, other non-oil, and other oil individuals that were not included in the Core12

allelic diversity present within the cultivated set, with a subset of $N = 192$ individuals capturing 79.6% of the total allelic diversity. Interestingly, the smallest subset ($N = 12$) captured nearly half of the allelic diversity (45.7%) present within the full set of 433 individuals.

## Discussion

### Genetic diversity

Knowledge of the amount and distribution of genetic diversity within the cultivated sunflower germplasm is of great value for ongoing crop improvement efforts. A variety of factors influence the amount and distribution of genetic diversity in crop gene pools, including the timing, duration, and extent of the domestication bottleneck, the number of independent domestication events, and the amount of gene exchange between domesticated lineages and their wild relatives following domestication (Doebley et al. 2006). While the details surrounding the origin of cultivated sunflower have been debated (e.g., Brown 2008; Lentz et al. 2008a, b; Heiser 2008; Rieseberg and Burke 2008; Smith 2008), the bulk of available evidence indicates that sunflower was likely domesticated ca. 4,000–5,000 years ago in east-central United States (Harter et al. 2004; Smith 2006). Moreover, both nuclear and chloroplast DNA evidence indicate that sunflower is likely the product of a

single origin of domestication (Harter et al. 2004; Wills and Burke 2006). While a number of crops are known to have undergone subsequent bottlenecks associated with breeding and improvement that have further reduced variation within subsets of their gene pool (e.g., Hyten et al. 2006), relatively little is known about the impact of such factors on levels and patterns of genetic variation across the sunflower gene pool [but see Burke et al. (2005), Chapman et al. (2008), and Fusari et al. (2008) for analyses of the diversity in a limited subset of the germplasm].

The results of this study indicate that the cultivated sunflower gene pool harbors approximately two-thirds (67%) of the SSR diversity present in a random, range-wide sample of its wild progenitor. This apparent loss of genetic diversity is likely due to the occurrence of a population bottleneck during domestication, which is a widespread phenomenon in crop species (Doebley et al. 2006). In maize, for example, an SSR survey revealed that the cultivar gene pool had roughly 88% of the gene diversity and 76% of the allelic diversity present in its wild progenitor, teosinte (Vigouroux et al. 2005). In soybean, Kuroda et al. (2010) found an even greater reduction of gene diversity, with the crop gene pool containing roughly 50% of the SSR diversity present in its wild progenitor, *Glycine soja*. Similarly, cultivated barley lines exhibit a 50% reduction in diversity at SSR loci when compared with their wild progenitor (Matus and Hayes 2002). Our results thus suggest that sunflower experienced a moderate domestication bottleneck, within the range of that which occurred in other crop species. Interestingly, there were no detectable differences in allelic diversity among the 12 different categories of sunflower, suggesting that the primary determinant of overall levels of genetic diversity across the sunflower gene pool was the initial domestication bottleneck. We cannot, however, rule out the possibility that gene flow and introgression during modern breeding efforts influenced the level of standing genetic variation within certain segments of the cultivated sunflower gene pool.

Observed heterozygosity within cultivated accessions was low, averaging less than 10% (vs. 29% in the wilds), similar to the amount of residual heterozygosity present within maize inbred lines (Vigouroux et al. 2008). This result is not particularly surprising given that cultivated sunflower has lost the sporophytic self-incompatibility that is characteristic of its wild progenitor and, as noted above, recent breeding efforts have largely focused on the development of inbred lines. Because wild sunflower is an obligate outcrosser, it seems unlikely that the relatively low level of observed versus expected heterozygosity is due to inbreeding (0.29 vs. 0.70, respectively). Rather, it seems more likely that this pattern is a byproduct of our sampling strategy, wherein we sampled a small number of individuals from a relatively large number of populations. This strategy was intended to capture as much diversity across the range of wild sunflower as possible, but has the potential to produce an apparent heterozygote deficit [i.e., Wahlund effect (Wahlund 1928; Halliburton 2004)].

## Population structure

Despite being fully interfertile and considered members of the same taxonomic species, wild and cultivated sunflower exhibit markedly different phenotypes with regard to traits like overall plant architecture (branching), self-incompatibility, seed size, and the size and number of flowering heads. The results of our STRUCTURE analysis were fully consistent with these observed phenotypic differences, as wild and cultivated sunflower were separated into genetically distinct clusters. In order to investigate population structure within the wild and cultivated subsets, we re-analyzed the two groups independently. The analysis of wild sunflower revealed the existence of four genetically distinct subgroups within our North American sample. The lack of a clear geographic signal amongst these groups suggests a history of extensive gene flow across the native range of wild sunflower. Historically, sunflower is thought to have been dispersed by both humans and other animals, including bison (Asch 1993). Moreover, wild sunflower is often found growing as a weed in disturbed habitats and thus seems likely to have been spread via anthropogenic forces.

Our results also revealed evidence of genetic substructure within the cultivated sunflower germplasm collection. More specifically, the STRUCTURE analysis revealed the presence of two genetically distinct groups. One of these was primarily composed of the RHA-oil lines, with the other group largely consisting of the remainder of the sampled collection. A distinction between R- and B- lines would be expected given their breeding history, which has involved the maintenance of somewhat distinct gene pools to maximize heterosis in hybrid crosses (Fehr 1987; Fick and Miller 1997). Interestingly, there was no such distinction within the non-oil lines, where R-lines grouped with B-lines, and both of those grouped with oil B-lines. This difference is likely due to differences in the breeding history and timing of development of the two types. More specifically, early hybrid sunflower breeding efforts focused mainly on high oil varieties, with attention later turning to the development of hybrid confectionery lines. Moreover, breeders have introgressed traits from oil lines into non-oilseed sunflower, including high oleic fatty acid content, which improves the shelf life of the achenes (Fick and Miller 1997). Significant $F_{ST}$ values also revealed genetic substructure within the cultivated germplasm with INRA-RHA and RHA-oil lines being the most different

from the other cultivar classifications. Despite the occurrence of significant population genetic structure between the RHA-oil lines and the remainder of the germplasm, however, the neighbor-joining tree based on allele sharing showed almost no discernible phylogenetic structure amongst accessions, suggesting that there are no deep divisions within the cultivated sunflower gene pool.

To further investigate the genetic composition of cultivated sunflower as it relates to wild sunflower, we statistically estimated the proportion of cultivars that had contributions from each of the four wild sunflower clusters. Interestingly, the majority of cultivated accessions, including the landraces, were estimated to have major contributions from wild cluster 3. As noted above, there is relatively little in the way of discernible geographic structure within and among the wild clusters, but it is noteworthy that cluster 3 (denoted in green in Fig. 1) is primarily composed of wild sunflower individuals from the east-central United States, where sunflower domestication is thought to have occurred (Heiser et al. 1969; Smith 1989; Rieseberg and Seiler 1990; Crites 1993; Harter et al. 2004; Smith 2006). The next largest contribution came from wild cluster 4 (denoted in purple in Fig. 1). This cluster also contains a number of wild sunflower individuals from populations in the east-central United States including all of the Arkansas, Iowa, Missouri, and Tennessee individuals surveyed here, though this cluster also contains a mixture of individuals from geographical disparate regions. This finding is again consistent with the proposed east-central North American origin of sunflower. Taken together, these results suggest that wild sunflower individuals, particularly those from clusters 1 and 2, may be a valuable source of unique genetic diversity for future breeding efforts.

## The nested core set and future directions

Beyond the basic biological insights provided by this work, our data also allowed for the identification of a nested core set of accessions that captures the maximum amount of allelic diversity for a given sample size. At the lowest level, 12 cultivars were sufficient to capture nearly 50% of the total allelic diversity present within the germplasm collection (105 of 230 alleles). These 12 individuals represented INRA, RHA, HA, oil, and non-oil accessions. The amount of allelic diversity present within the nested core sets increased steadily with increasing sample size, reaching nearly 90% allelic representation at 288 individuals. This core collection, which we have advanced via single seed descent, represents an invaluable resource for sunflower researchers, particularly with regard to future association mapping efforts.

Association mapping, also known as linkage disequilibrium (LD) mapping, involves correlating molecular variation with phenotypic variation across a diverse collection of genotypes. Due to the large amount of historical recombination that is captured within an association mapping population, this approach typically provides much higher resolution than is available with traditional QTL mapping approaches (Pritchard et al. 2000b; Purcell et al. 2003). Given that LD is known to decay relatively rapidly in sunflower (Liu and Burke 2006; Kolkman et al. 2007; Fusari et al. 2008), association mapping has the potential to provide gene-level resolution for the mapping of functional variation in sunflower. A major obstacle for association mapping studies is that population structure can lead to false associations when the frequency of a phenotype varies across subpopulations (Pritchard et al. 2000a; Buckler and Thornsberry 2002). However, statistical methods have been developed to control for population structure and kinship based on genotypic data from so-called "background markers" (such as those employed here), thereby reducing the likelihood of spurious associations (e.g., Yu et al. 2006). While our data revealed the presence of population structure within the cultivated sunflower gene pool, we found no evidence for deep genetic divisions within the germplasm collection. When combined with the observed levels of gene diversity and the apparently rapid decay of linkage disequilibrium across the genome, association mapping appears to be a promising approach for the genetic dissection of complex traits in sunflower.

**Acknowledgments** We thank Patrick Vincourt from INRA as well as the staff of the USDA North Central Regional Plant Introduction Station for providing us with germplasm. Patrick Vincourt also provided valuable scientific input. We thank John Hvala, Jason Strever, Michael Payne, and the University of Georgia greenhouse staff for assistance with data collection and plant maintenance. We are grateful to members of the Burke lab for useful comments on an earlier version of this manuscript. Finally, we would like to thank Brent Hulke of the USDA for helpful discussions regarding this work. This work was funded by a grant from the USDA National Institute of Food and Agriculture to JMB (2008-35300-19263).

## References

Acquaah G (2006) Principles of plant genetics and breeding. Blackwell, Oxford

Asch DL (1993) Common sunflower (*Helianthus annuus* L.): the pathway toward domestication. In: Proceedings of the 58th Annual Meeting of the Society for American Archaeology. Society for American Archaeology, St. Louis, pp 1–15

Beard B (1982) Registration of *Helianthus* germplasm pools III and IV. Crop Sci Crop Sci 22:1276–1277

Bolker B, Okuyama T, Bjorndal K, Bolten A (2003) Sea turtle stock estimation using genetic markers: accounting for sampling error of rare genotypes. Ecol App 12:763–775

Brothers ME, JF Miller (1999) Core subset for the cultivated sunflower collection. In: Proceedings of 21st Sunflower Research Workshop. Fargo, ND, pp 124–127

Brown C (2008) A lack of linguistic evidence for domesticated sunflower in pre-Columbian Mesoamerica. Proc Natl Acad Sci USA 105:E47

Buckler ES, Thornsberry J (2002) Plant molecular diversity and applications to genomics. Curr Opin Plant Biol 5:107–111

Burke JM, Knapp SJ, Rieseberg LH (2005) Genetic consequences of selection during the evolution of cultivated sunflower. Genet 171:1933–1940

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* L.). Plant Cell 20:2931–2945

Cheres MT, Knapp SJ (1998) Ancestral origins and genetic diversity of cultivated sunflower: coancestry analysis of public germplasm. Crop Sci 38:1476–1482

Crites GD (1993) Domesticated sunflower in fifth millennium B.P. Temporal context: new evidence from Middle Tennessee. Am Antiq 58:146–148

Cronn R, Brothers M, Klier K, Bretting PK, Wendel JF (1997) Allozyme variation in domesticated annual sunflower and its wild relatives. Theor Appl Genet 95:532–545

Doebley JF, Gaut BS, Smith BD (2006) The molecular genetics of crop domestication. Cell 127:1309–1321

Dong GJ, Liu GS, Li KF (2007) Studying genetic diversity in the core germplasm of confectionary sunflower (*Helianthus annuus* L.) in China based on AFLP and morphological analysis. Russ Jour Genet 43:627–635

Doyle JL, Doyle JJ (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem 19:11–15

Evanno, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–2620

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction sites. Genet 131:479–491

Fehr WR (1987) Principles of cultivar development, vol 1, theory and technique. Macmillan, New York

Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Department of Genome Sciences, University of Washington, Seattle

Fick GN, Miller JF (1997) Sunflower breeding. In: Scheiter AA (ed) Sunflower production and technology. Amer Society Agron, Madison, pp 395–440

Fusari CM, Lia VV, Hopp HE, Heinz RA, Paniego NB (2008) Identification of single nucleotide polymorphisms and analysis of sinkage disequilibrium in sunflower elite inbred lines using the candidate gene approach. BMC Plant Bio 8:7

Halliburton R (2004) Introduction to population genetics. Pearson-Prentice Hall, New Jersey

Harter AV, Gardner KA, Falush D, Lentz DL, Bye RA, Rieseberg LH (2004) Origin of extant domesticated sunflowers in eastern North America. Nat 430:201–205

Heiser CB (2008) How old is the sunflower in Mexico? Proc Natl Acad Sci USA 105:E48

Heiser CB, Smith DM, Clevenger S, Martin WC (1969) The North American sunflowers (*Helianthus*). Mem Torrey Bot Club 22:1–218

Hongtrakul V, Huestis GM, Knapp SJ (1997) Amplified fragment length polymorphisms as a tool for DNA fingerprinting sunflower germplasms: genetic diversity among oilseed inbred lines. Theor Appl Genet 95:400–407

Hurlbert SH (1971) The non-concept of species diversity: a critique and alternative parameters. Ecol 52:577–586

Hyten DL, Song Q, Zhu Y, Choi I-Y, Nelson RL, Costa JM et al (2006) Impacts of genetic bottlenecks on soybean genome diversity. Proc Natl Acad Sci USA 103:16666–16671

Kalinowski ST (2004) Counting alleles with rarefaction: private alleles and hierarchical sampling designs. Conserv Genet 5:539–543

Kalinowski ST (2005) HP-Rare: a computer program for performing rarefaction on measures of allelic diversity. Mol Ecol Notes 5:187–189

Kinman ML (1970) New developments in the USDA and state experiment station breeding programs. In: Proceedings of 4th International Sunflower Conference, Int Sunflower Assoc, Paris, pp 181–183

Kolkman JM, Berry ST, Leon AJ, Slabaugh MB, Tang S, Gao W et al (2007) Single nucleotide polymorphisms and linkage disequilibrium in sunflower. Genet 177:457–468

Korell M, Mosges G, Friedt W (1992) Construction of a sunflower pedigree map. Helia 15:7–16

Kuroda Y, Kaga A, Tomooka N, Vaughan D (2010) The origin and fate of morphological intermediates between wild and cultivated soybeans in their natural habitats in Japan. Mol Ecol 19:2346–2360

Leclercq P (1969) Une Sterilite male cytoplasmique chez le tournesol. Ann Amélior Plant 19:99–106

Lentz DL, Pohl MD, Alvarado JL, Tarighat S, Bye R (2008a) Sunflower (*Helianthus annuus* L.) as a pre-Columbian domesticate in Mexico. Proc Natl Acad Sci USA 105:6232–6237

Lentz DL, Pohl MD, Bye R (2008b) Reply to Rieseberg and Burke, Heiser, Brown, and Smith: Molecular, linguistic, and archaeological evidence for domesticated sunflower in pre-Columbian Mesoamerica. Proc Natl Acad Sci USA 105:E49–E50

Liu A, Burke JM (2006) Patterns of nucleotide diversity in wild and cultivated sunflower. Genet 173:321–330

Liu K, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinforma 21:2128–2129

Matus IA, Hayes PM (2002) Genetic diversity in three groups of barley germplasm assessed by simple sequence repeats. Genome 45:1095–1106

Morariu VI, Srinivasan BV, Raykar VC, Duraiswami R, Davis LS (2008) Automatic online tuning for fast Gaussian summation. In: Koller D, Schuurmans D, Bengio Y, Bottou L (eds) Adv Neural Inf Process Sys (NIPS) 1113–1120

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. Genet 89:583–590

Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Mol Ecol Notes 6:288–295

Pella JJ, Milner GB (1987) Use of genetic markers in stock composition analysis. In: Ryman N, Utter FW (eds) Population genetics and fisheries management. UW Press, Seattle, pp 247–276

Petit RJ, El Mousadik A, Pons O (1998) Identifying populations for conservation on the basis of genetic markers. Conserv Biol 12:844–855

Pritchard JK, Stephens M, Donnelly P (2000a) Inference of population structure using multilocus genotype data. Genet 155:945–956

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000b) Association mapping in structured populations. Am J Hum Genet 67:170–181

Purcell S, Cherny SS, Sham PC (2003) Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. Bioinforma 19:149–150

Pustovoit VS (1964) Conclusions of work on the selection and seed production of sunflowers (in Russian). Agrobio 5:662–697

Putt ED (1997) Early history of sunflower. In: Scheiter AA (ed) Sunflower production and technology. Amer Society Agron, Madison, pp 1–19

Rieseberg LH, Burke JM (2008) Molecular evidence and the origin of the domesticated sunflower. Proc Natl Acad Sci USA 105:E46

Rieseberg LH, Seiler GJ (1990) Molecular evidence and the origin and development of the domesticated sunflower (*Helianthus annum* Asteraceae). Econ Bot 44:79–91

Robertson JA, Burns EE (1975) Use of sunflower seed in food products. Crit Rev Food Sci 6:201–240

Robertson JA, Morrison WH (1977) Effect of heat and frying on sunflower oil stability. Symposium: Oilseeds-new foods for tomorrow. A Oil Chem Soc, New Orleans, LA pp 77A–81A

Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. Nat Biotechnol 18:233–234

Smith BD (1989) Origins of agriculture in eastern North America. Science 246:1566–1571

Smith BD (2006) Eastern North America as an independent center of plant domestication. Proc Natl Acad Sci USA 103:2223–12228

Smith BD (2008) Winnowing the archaeological evidence for domesticated sunflower in pre-Columbian Mesoamerica. Proc Natl Acad Sci USA 105:E45

Smouse PE, Waples RS, Tworek JA (1990) A genetic mixture analysis for use with incomplete source population data. Can J Fish Aquat Sci 47:620–634

Soleri D, Cleveland D (1993) Hopi crop diversity and change. J Ethnobiol 13:203–231

Tanksley, Nelson JC (1996) Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. Theor Appl Genet 92:191–203

USDA, 2006. Sunflower Inbred Lines. http://www.ag.ndsu.nodak.edu/aginfo/seedstock/varieties/VH-SUNF.htm

Vavilov NI (1940) The new systematics of cultivated plants. In: Huxley J (ed) The new systematics. Clarendon Press, Oxford, pp 549–566

Vigouroux Y, Mitchell S, Matsuoka Y, Hamblin M, Kresovich S, Smith JSC, Jaqueth J, Smith OS, Doebley J (2005) An analysis of genetic diversity across the maize genome using microsatellites. Genet 169:1617–1630

Vigouroux Y, Glaubitz J, Matsuoka Y, Goodman MM, Sanchez GJ, Doebley J (2008) Population structure and genetic diversity of New World maize races assessed by DNA microsatellites. Am J Bot 95:1240–1253

Wahlund S (1928) Zusammensetzung von Population und Korrelationserscheinung vom Standpunkt der Vererbungslehre aus betrachtet. Hereditas 11:65–106

Wills DM, Burke JM (2006) Chloroplast DNA variation confirms a single origin of domesticated sunflower (*Helianthus annuus* L.). J Hered 97:403–408

Wills DM, Hester ML, Liu A, Burke JM (2005) Chloroplast SSR polymorphisms in the Compositae and the mode of organellar inheritance in *Helianthus annuus*. Theor Appl Genet 110:941–947

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

Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

Yue B, Cai WX, Vick BA, Hu JG (2009) Genetic diversity and relationships among 177 public sunflower inbred lines assessed by TRAP markers. Crop Sci 49:1242–1249