DOI: 10.1534/genetics.107.074054

Single Nucleotide Polymorphisms and Linkage Disequilibrium in Sunflower

Judith M. Kolkman,*,[1] Simon T. Berry,[†] Alberto J. Leon,[‡] Mary B. Slabaugh,* Shunxue Tang,*,[§] Wenxiang Gao,[§] David K. Shintani, John M. Burke[††] and Steven J. Knapp*,[§,2]**

**Department of Crop and Soil Science, Oregon State University, Corvallis, Oregon 97331, [†]Advanta Seeds UK, Norfolk, PE31 8LS, United Kingdom, [‡]Advanta Semillas, Balcarce Research Station, Argentina, [§]Center for Applied Genetic Technologies, The University of Georgia, Athens, Georgia 30602, **Department of Biochemistry and Molecular Biology, The University of Nevada, Reno, Nevada 89557 and [††]Department of Plant Biology, The University of Georgia, Athens, Georgia 30602*

Manuscript received March 31, 2007
Accepted for publication July 13, 2007

ABSTRACT

Genetic diversity in modern sunflower (*Helianthus annuus* L.) cultivars (elite oilseed inbred lines) has been shaped by domestication and breeding bottlenecks and wild and exotic allele introgression—the former narrowing and the latter broadening genetic diversity. To assess single nucleotide polymorphism (SNP) frequencies, nucleotide diversity, and linkage disequilibrium (LD) in modern cultivars, alleles were resequenced from 81 genic loci distributed throughout the sunflower genome. DNA polymorphisms were abundant; 1078 SNPs (1/45.7 bp) and 178 insertions-deletions (INDELs) (1/277.0 bp) were identified in 49.4 kbp of DNA/genotype. SNPs were twofold more frequent in noncoding (1/32.1 bp) than coding (1/62.8 bp) sequences. Nucleotide diversity was only slightly lower in inbred lines ($\theta = 0.0094$) than wild populations ($\theta = 0.0128$). Mean haplotype diversity was 0.74. When extraploted across the genome (~3500 Mbp), sunflower was predicted to harbor at least 76.4 million common SNPs among modern cultivar alleles. LD decayed more slowly in inbred lines than wild populations (mean LD declined to 0.32 by 5.5 kbp in the former, the maximum physical distance surveyed), a difference attributed to domestication and breeding bottlenecks. SNP frequencies and LD decay are sufficient in modern sunflower cultivars for very high-density genetic mapping and high-resolution association mapping.

TECHNOLOGICAL advances in DNA sequencing have facilitated direct analyses of nucleotide diversity and large-scale single nucleotide polymorphism (SNP) discovery in diverse eukaryotes, as well as the development of highly parallel SNP genotyping methods and high-resolution linkage disequilibrium (LD)-based association mapping approaches for identifying functionally important nucleotide polymorphisms (Jorde 1995, 2000; Lindblad-Toh *et al.* 2000; Risch 2000; Syvanen 2001, 2005; Buckler and Thornsberry 2002; Nordborg and Tavare 2002; Flint-Garcia *et al.* 2003; Weigel and Nordborg 2005; Kim *et al.* 2006). Very high DNA marker densities are needed for identifying DNA polymorphisms linked to phenotypic and quantitative trait loci through whole-genome association mapping approaches and can only be achieved using SNPs, the most abundant class of DNA polymorphisms (Collins *et al.* 1998; Aquadro *et al.* 2001; Wiltshire *et al.* 2003). While simple sequence repeat (SSR) and insertion-deletion (INDEL) markers are versatile and highly portable, and have been mainstays in molecular breeding and genomics applications (Taramino and Tingey 1996; Bhattramakki *et al.* 2002), SNPs are significantly more common than either and critical for massively parallel array-facilitated genotyping (Lindblad-Toh *et al.* 2000; Syvanen 2001, 2005; Buckler and Thornsberry 2002; Rafalski 2002a,b).

SNP abundance and LD decay are highly variable in eukaryotic genomes and affected by natural, domestication, and breeding history, mating systems, mutation, migration, genomic rearrangements, recombination, and other factors (Chapman and Thompson 2001; Hudson 2001; Buckler and Thornsberry 2002; Stumpf 2002; Greenwood *et al.* 2004; Rafalski and Morgante 2004). Typically, SNPs are less abundant, and LD decays more slowly in autogamous than allogamous species, domesticated than wild genotypes, and inbred than outbred genotypes (Ching *et al.* 2002; Nordberg *et al.* 2002; Nordberg and Tavare 2002; Flint-Garcia *et al.* 2003; Rafalski and Morgante 2004; Shifman *et al.*

Sequence data from this article have been deposited with EMBL/GenBank Data Libraries under accession nos. EF469860-EF469941 and EF460879-EF462190.

[1]*Present address:* Department of Plant Pathology, Cornell University, Ithaca, NY 14853.

[2]*Corresponding author:* Center for Applied Genetic Technologies, 111 Riverbend Road, The University of Georgia, Athens, GA 30602. E-mail: sjknapp@uga.edu.

2003; Ingvarsson 2005). For example, SNPs are significantly more frequent in maize (*Zea mays* L.; 1 SNP/61 bp), a predominantly allogamous species, than soybean (*Glycine max* L.; 1 SNP/273 bp to 1 SNP/343 bp), a predominantly autogamous species (Remington *et al.* 2001; Tenaillon *et al.* 2001, 2002; Ching *et al.* 2002; Zhu *et al.* 2003; Van *et al.* 2005). Moreover, LD decays more slowly (persists over much longer tracts of DNA) in soybean (>50 kbp) than maize (400–1500 bp). LD decayed more rapidly in exotic outbred germplasm than elite inbred lines in maize, a difference attributed to the effects inbreeding and selection (Ching *et al.* 2002; Rafalski and Morgante 2004). The persistence of LD decreases the density of DNA marker loci needed for identifying phenotypic–genotypic associations, but decreases resolution (Cardon and Bell 2001; Cardon and Abecasis 2003; Rafalski and Morgante 2004).

Sunflower (*Helianthus annuus* L.), a predominantly allogamous species, should display patterns of nucleotide diversity and LD similar to maize and other allogamous species. Genetic diversity in modern sunflower cultivars (elite oilseed inbred lines and hybrids) has been shaped by domestication and breeding, as well as the introgression of alleles from wild and exotic germplasm (migration) (Cheres and Knapp 1998; Tang and Knapp 2003; Harter *et al.* 2004; Burke *et al.* 2005). Domestication and breeding create population bottlenecks, decrease genetic diversity, and increase LD, whereas migration increases genetic diversity and decreases LD (Ching *et al.* 2002; Rafalski and Morgante 2004). The abundance and distribution of SNPs in elite oilseed inbred line alleles has only been reported for a few genic loci in sunflower (Kolkman *et al.* 2004; Hass *et al.* 2006; Schuppert *et al.* 2006; Tang *et al.* 2006b), and LD has only been surveyed in Native American land races and other exotic cultivars and wild populations (Liu and Burke 2006). Kolkman *et al.* (2004) found significant differences in SNP frequencies among acetohydroxyacid synthase alleles resequenced from inbred lines and wild populations, a pattern predicted from analyses of SSR diversity (Tang and Knapp 2003). Liu and Burke (2006) surveyed nucleotide diversity and LD in nine genic loci in wild populations and exotic germplasm accessions (Native American land races and prehybrid era open-pollinated confectionery and oilseed cultivars); only one elite inbred line allele (HA89) was resequenced. SNPs were twofold more abundant in wild populations (1 SNP/19.9 bp) than exotic germplasm accessions (1 SNP/38.8 bp), exotic alleles harbored half of the nucleotide diversity found in wild alleles, and LD decayed within ~200 bp in wild alleles and ~1100 bp in exotic alleles. Here, we report SNP frequencies, nucleotide diversity, and LD in elite sunflower inbred lines alleles resequenced from 82 previously mapped restriction fragment length polymorphism (RFLP) marker loci distributed throughout the sunflower genome ($2n = 2x = 34$) (Berry *et al.* 1995; Gedil *et al.* 2001; Yu *et al.* 2002, 2003).

MATERIALS AND METHODS

Plant materials and allele resequencing: DNA polymorphisms were surveyed in two wild (ANN1238 and ANN1811) and 10 elite inbred line alleles resequenced from 82 previously mapped RFLP marker loci (ZVG1-ZVG17, ZVG19-ZVG81, ZVG152, and ZVG668) (Tables 1 and 2; supplemental Table 1 at http://www.genetics.org/supplemental/) (Berry *et al.* 1994, 1995). We sequenced a single phase known allele/resequenced amplicon (RSA) from each genotype by cloning genomic DNA amplicons and randomly selecting and sequencing a single clone/RSA/genotype; one or two DNA fragments (amplicons) were resequenced per RFLP locus. Leaves were harvested from 10 4- to 6-wk-old plants from each germplasm accession and bulked. Genomic DNA samples were isolated from each bulk using a modified CTAB method (Murray and Thompson 1980). Of the 82 RFLP probes, 78 were cDNA clones developed from RNAs isolated from etiolated seedlings and 4 were *Pst*I-digested genomic DNA clones (ZVG9, ZVG16, ZVG19, and ZVG51) (Berry *et al.* 1994). The probe inserts were sequenced, and BLASTX analyses of the sequences were performed against the National Center for Biotechnology Information (NCBI) Protein Database (http://www.ncbi.nlm.nih.gov) to identify putative functions using a probability threshold of $\leq e^{-15}$ (Altschul *et al.* 1990; Altschul and Gish 1996; Mcginnis and Madden 2004). The probe insert sequences were used as templates for designing resequencing primers using Primer3 (http://frodo.wi.mit.edu) and manual selection. Forward and reverse primer sites were chosen as close as possible to opposite ends of the reference allele sequences so as to amplify the longest DNA fragments possible from each locus (supplemental Table 1). Genomic DNA fragments were amplified using long-distance PCR (LD-PCR) (Barnes 1994) in most cases and PCR in a few cases. PCRs and LD-PCRs were performed by adding 30–60 ng of genomic DNA to a 20-μl PCR mix containing 1× buffer, 2 mm $MgSO_4$, 0.3 mm dNTPs, 0.3 μm of forward and reverse primers, 0.5 U of Platinum *Taq* DNA Polymerase High Fidelity (Invitrogen, Carlsbad, CA), and dH_2O to a final volume of 20 μl. For LD-PCR, genomic DNAs were amplified using one cycle at 94° for 4 min, followed by 10 cycles at 94° for 10 sec, 58° for 1 min, and 68° for up to 12 min (1 min per kb), 25 cycles at 94° for 10 sec, 58° for 1 min, and 68° for up to 12 min plus 10 sec per cycle, and one cycle at 72° for 20 min; annealing temperatures ranged from 55° to 62°, and extension times ranged from 2 to 12 min. Genomic DNA amplicons were cloned using the Invitrogen TOPO TA-cloning method. We selected and single-pass sequenced a single clone for each genotype by amplicon combination from one or both ends at the University of Nevada, Reno Genomics Center on an Applied Biosystems Prism 3730 DNA Sequencer (Foster City, CA). By sequencing a single cloned amplicon, we acquired a single phase known allele from each genotype.

DNA sequence analyses: DNA sequences were aligned using Contig Express and AlignX (Vector NTI; Invitrogen), low quality base calls (<PHRED 20) were trimmed using PHRED (Ewing and Green 1998; Ewing *et al.* 1998), and trimmed allele sequence alignments were used for nucleotide diversity analyses. Polymorphic sites and synonymous and nonsynonymous SNPs were identified and counted using DnaSP (Rozas and Rozas 1999; Rozas *et al.* 2003) (http://www.ub.es/dnasp/). DNA sequence alignments were visually inspected to identify and count polymorphisms. Nucleotide diversity statistics (π and θ) were estimated for synonymous, nonsynonymous, and silent (synonymous and noncoding) sites, where π is the mean number of nucleotide differences per site between two allele sequences (Nei 1987), and θ is the mean number of segregating

sites (Watterson 1975; Halushka *et al.* 1999). Haplotype diversity was estimated as described by Nei (1987).

LD analyses were performed on RSAs >1 kbp in length harboring at least 10 polymorphic sites. The physical distances separating pairs of polymorphic sites between independent RSAs amplified from opposite ends of a locus were estimated from DNA fragment length estimates (supplemental Table 1 at http://www.genetics.org/supplemental/). Using DnaSP, we estimated the minimum number of recombination events (RM) in inbred line alleles using the four-gamete test (Hudson and Kaplan 1985), proportion of adjacent polymorphisms in perfect disequilibrium (*B*) (Wall 1999), and strength of LD between pairs of polymorphic sites (estimated as the squared allele frequency correlation, r^2) (Weir 1996). The decay of LD against physical distance was modeled using nonlinear regression methods described by Remington *et al.* (2001). Briefly, SAS PROC NLIN (Cary, NC) was used to fit r^2 estimates (pooled across loci) to a model of the expected level of r^2 at drift-recombination equilibrium, allowing for a low level of mutation and finite sample size (see Appendix 2 of Hill and Weir 1988). Although factors such as the nonindependence of linked sites and nonequilibrium populations can reduce the precision of such analyses and introduce bias, they are still useful for investigating the overall rate of decay of LD (see Ingvarsson 2005).

RESULTS

Allele resequencing and putative functions of the resequenced loci: The inserts of 4 genomic DNA and 78 cDNA clones previously used as RFLP probes (Berry *et al.* 1994, 1995; Gedil *et al.* 2001) were sequenced and ranged in length from 97 to 3025 bp (supplemental Table 1 at http://www.genetics.org/supplemental/; GenBank accession nos. EF469860–EF469941). The putative functions of 48 of the 82 loci were inferred from BLASTX searches (Table 1). For the other 34 loci, BLASTX searches either failed to identify proteins (probabilities were $>e^{-15}$) or identified unknown proteins.

The 82 RFLP markers were known to be low-copy and polymorphic among elite inbred lines (Berry *et al.* 1994, 1995). The primer pairs selected for allele resequencing produced amplicons ranging in length from 97 to ~10,000 bp across loci (Figure 1; supplemental Table 1 at http://www.genetics.org/supplemental/). Of the 82 primer pairs, 77 produced paralog-specific amplicons and 31 of the 77 spanned introns, INDELs, or both (amplicon lengths are shown in supplemental Table 1). By sequencing cloned amplicons, a single phase-known allele was resequenced from each genotype (Table 2). Collectively, 1312 RSAs and 129 DNA sequence alignments were produced for 81 of the 82 loci; allele sequences could not be produced for the ZVG46 locus (GenBank accession nos. EF469941–EF462190; allele sequence alignments are displayed in supplemental Figure 1). Nucleotide polymorphisms were surveyed in 84 to 100 DNA sequences/genotype and 49.4 kbp of DNA sequence/genotype (Table 3). Nucleotide diversity analyses were performed on 107 DNA sequence alignments comprised of 6 to 10 inbred line allele sequences each from 71 of the 81 resequenced loci. The other 22 DNA sequence alignments were either comprised of 6 or fewer inbred line allele sequences, paralogous RSAs, or both specific and nonspecific RSAs.

Nucleotide diversity: SNPs were identified in every locus, although two RSAs (ZVG5-F and ZVG33) only had one SNP each, and one RSA (ZVG64) lacked DNA polymorphisms among inbred line alleles (supplemental Figure 1 at http://www.genetics.org/supplemental/). DNA polymorphisms were abundant among inbred line alleles; 1078 SNPs (1/45.7 bp) and 178 INDELs (1/277.0 bp) were identified in the 49.4 kbp of DNA sequence surveyed (Table 3). Of the 1078 SNPs, 55.9% were transitions and 44.1% were transversions. SNPs were twofold more frequent in noncoding (1/32.1 bp) than coding (1/62.8 bp) sequences, most frequent in introns (1/29.2 bp), and second most frequent in UTRs (1/37.5 to 1/42.3 bp). Synonymous SNPs (1/139.5 bp) were sixfold more frequent than nonsynonymous SNPs (1/22.5 bp).

The mean number of segregating sites was $\theta = 0.0094$, and the mean number of pairwise sequence differences was $\pi = 0.0107$ among RSAs (Table 3). Nucleotide diversity was twofold greater in noncoding than coding sequences, sixfold greater for SNPs ($\pi = 0.0092$) than INDELs ($\pi = 0.0016$), and greatest in introns ($\pi = 0.01480$). Nonsynonymous substitutions ($\pi_{nonsyn} = 0.0028$) were sixfold less prevalent than synonymous substitutions ($\pi_{syn} = 0.0176$), suggesting variability among loci has primarily been produced by purifying selection (Figure 2; Table 3). π_{nonsyn} ranged from 0.0 to 0.055, and π_{syn} ranged from 0.0 to 0.109 among RSAs (nucleotide diversity statistics for individual RSAs are shown in supplemental Table 2 at http://www.genetics.org/supplemental/). Only two RSAs had π_{nonsyn}/π_{syn} ratios >1.0 ($\pi_{nonsyn}/\pi_{syn} = 1.12$ for ZVG47 and 1.10 for ZVG80-R) (supplemental Table 2).

θ_{silent} ranged from 0.0008 to 0.109 among RSAs, a 136-fold difference (Figure 3; supplemental Table 2 at http://www.genetics.org/supplemental/). RSAs on two linkage groups (6 and 15), as a whole, had significantly fewer silent substitutions than RSAs on the other 15 linkage groups. θ_{silent} ranged from 0.0029 for ZVG28-F to 0.0113 for ZVG27 on linkage group (LG) 6 and from 0.0022 for ZVG69-R to 0.0147 for ZVG70-R on LG 15.

SNP allele frequencies, heterozygosities, and haplotype diversity: The mean frequency of the less common SNP allele (f_r) was 0.31 amongst 1078 SNPs, and the SNP heterozygosity (h_s) mean was 0.41 amongst the 10 inbred lines, only 0.09 less than the theoretical maximum (0.50) for a biallelic DNA marker (Figure 4). f_r ranged from 0.17 to 0.50, h_s ranged from 0.28 to 0.50, and the f_r and h_s distributions were nearly uniform. Both distributions were left truncated because singleton SNPs ($f_r \leq 0.125$) were not counted so as to minimize false positives (sequencing errors) and avoid upwardly biasing SNP frequencies and downwardly biased SNP heterozygosities.

TABLE 1

Putative functions of resequenced sunflower loci inferred by BLASTX

RFLP marker locus[a]	Probe sequence length (bp)	GenBank accession no.	Putative function
ZVG1	600	EF469860	Betacyclase
ZVG2	1142	EF469861	Oxidoreductase
ZVG3	332	EF469862	Ubiquinol-cytochrome-c reductase
ZVG4	917	EF469863	Elongation factor
ZVG5	2194	EF469864	—
ZVG6	1134	EF469865	S-adenosylmethionine synthetase
ZVG7	1671	EF469866	Lipoxygenase
ZVG8	1510	EF469867	—
ZVG9	524	EF469868	—
ZVG10	805	EF469869	1-Deoxy-D-xylulose-5-phosphate synthase
ZVG11	1037	EF469870	40S Ribosomal protein S3A
ZVG12	1425	EF469871	Phosphate transporter
ZVG13	1465	EF469872	DNA binding/transcription factor
ZVG14	1751	EF469873	RelA-SpoT like protein
ZVG15	2368	EF469874	Protein-binding/transcription regulator
ZVG16	1677	EF469875	—
ZVG17	611	EF469876	—
ZVG668	760	EF469877	—
ZVG19	603	EF469878	—
ZVG20	1615	EF469879	Beta-tubulin
ZVG21	229	EF469880	—
ZVG22	596	EF469881	—
ZVG23	2464	EF469882	Peroxisomal targeting signal 1 receptor
ZVG24	1402	EF469883	Ribosomal protein L3
ZVG25	637	EF469884	—
ZVG26	716	EF469885	Shaggy-related protein kinase
ZVG27	963	EF469886	DNA binding/transcription factor
ZVG28	1599	EF469887	Ca^{2+}/H+ exchanger
ZVG29	172	EF469888	—
ZVG30	1578	EF469889	Protein kinase
ZVG31	688	EF469890	Heat shock protein
ZVG32	721	EF469891	NAD-dependent sorbitol dehydrogenase
ZVG33	97	EF469892	—
ZVG34	1963	EF469893	Gibberellin response modulator
ZVG35	435	EF469894	ABC transporter
ZVG36	818	EF469895	Photosystem I subunit
ZVG37	2120	EF469896	Heat shock protein hsp70
ZVG38	661	EF469897	—
ZVG39	2762	EF469898	Beta-galactosidase
ZVG40	672	EF469899	Ribosomal protein S19
ZVG41	1542	EF469900	—
ZVG42	664	EF469901	Developmental protein
ZVG43	1442	EF469902	UDP-GlcNAc:dolichol phosphate N-acetylglucosamine-1-phosphate transferase
ZVG44	873	EF469903	Ribosomal protein
ZVG45	517	EF469904	Ribosomal protein
ZVG46	1164	EF469905	Ribulose bisphosphate carboxylase/oxygenase activase (RuBisCO activase)
ZVG47	585	EF469906	Protein kinase
ZVG48	2367	EF469907	—
ZVG49	1021	EF469908	Hydrolase
ZVG50	352	EF469909	Ribulose bisphosphate carboxylase small chain chloroplast precursor
ZVG51	778	EF469910	—
ZVG52	1982	EF469911	—
ZVG53	918	EF469912	Tonoplast intrinsic protein
ZVG54	431	EF469913	Initiation factor eIF4A-15

(*continued*)

TABLE 1

(Continued)

RFLP marker locus[a]	Probe sequence length (bp)	GenBank accession no.	Putative function
ZVG55	2576	EF469914	F-box protein
ZVG56	1819	EF469915	Fiddlehead-like protein
ZVG57	643	EF469916	—
ZVG58	191	EF469917	—
ZVG59	1456	EF469918	—
ZVG60	1271	EF469919	—
ZVG61	367	EF469920	GTP-binding protein
ZVG62	1855	EF469921	Calcium ion binding/peptidase
ZVG63	1681	EF469922	—
ZVG64	701	EF469923	CDC-48-like protein
ZVG65	784	EF469924	—
ZVG66	509	EF469925	Ribosomal protein S4
ZVG67	375	EF469926	—
ZVG68	1003	EF469927	—
ZVG69	3025	EF469928	—
ZVG70	828	EF469929	Ribulose-1,5-bisphosphate carboxylase/oxygenase activase
ZVG71	904	EF469930	—
ZVG72	695	EF469931	—
ZVG73	546	EF469932	—
ZVG74	1936	EF469933	—
ZVG75	1286	EF469934	Glutamine synthetase
ZVG76	404	EF469935	—
ZVG77	1042	EF469936	Beta-amylase
ZVG78	817	EF469937	—
ZVG79	934	EF469938	Subunit of oxygen evolving system of photosystem II
ZVG80	1232	EF469939	—
ZVG81	1540	EF469940	Protein kinase
ZVG152	491	EF469941	—

[a] The RFLP marker loci were previously mapped using 4 genomic DNA and 78 cDNA probes (Berry *et al.* 1994, 1995; Gedil *et al.* 2001).

Haplotype diversities (h_d) ranged from 0.36 to 1.00, and mean haplotype diversity was 0.74 among inbred line and wild alleles (Figure 3; supplemental Figure 2 at http://www.genetics.org/supplemental/). The probability of observing one or more SNPs between two elite inbred line alleles drawn at random with replacement from the resequenced inbred line alleles (p_s) was 0.448 among cytoplasmic-genic (CMS) fertility maintainer (B) lines, 0.449 among CMS fertility restorer (R) lines, and 0.569 among B- and R lines. The number of haplotypes/locus ranged from 1 to 9 among 10 inbred line alleles and 2 to 11 among 12 inbred line and wild alleles (Table 2; numerical haplotypes are diplayed for each RSA in supplemental Figure 2). The mean number of haplotypes/locus was 2.3 among R-, 2.4 among B-, and 3.7 among B- and R-line alleles. The percentage of unique haplotypes ranged from 10.9 for ZENB13 to 27.4 for RHA373 among inbred line alleles and from 68.4 to 70.8 for the 2 wild alleles (Figure 5).

LD: LD statistics were estimated for 30 loci satisfying the criteria necessary for inclusion in our analyses (Figure 6). While LD varied across loci, with *B* (Wall

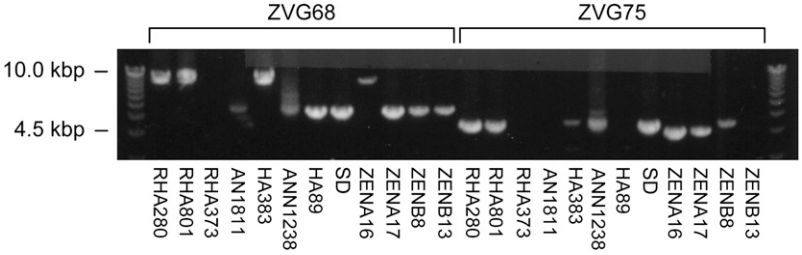

Figure 1.—Genomic DNA fragments for two genic loci (ZVG68 and ZVG75) amplified from 10 elite inbred lines and two wild populations (ANN1811 and AN1238).

TABLE 2

Sunflower inbred lines and wild populations selected for allele resequencing

Germplasm accession[a]	Plant introduction number	Germplasm type[b]
RHA280	PI 552943	Confectionery R-line
RHA801	PI 599768	Oilseed R-line
RHA373	PI 560141	Oilseed R-line
ZENA16	—	Oilseed R-line
ZENA17	—	Oilseed R-line
HA89	PI 599773	Oilseed B-line
HA383	PI 578872	Oilseed B-line
SD	PI 413039	Oilseed B-line
ZENB8	—	Oilseed B-line
ZENB13	—	Oilseed B-line
ANN1811	PI 494567	Wild population (Texas)
ANN1238	—	Wild population (Nebraska)

[a] Seeds of public inbred lines, identified by plant introduction (PI) numbers, were supplied by the United States Department of Agriculture (USDA) Agricultural Research Service (ARS) National Plant Germplasm System (http://www.ars-grin.gov/npgs/) or the USDA-ARS Northern Crop Science Research Laboratory. Seeds of proprietary inbred lines (ZENA16, ZENA17, ZENB8, and ZENB13) were supplied by Advanta Semillas, Balcarce, Argentina.

[b] B lines are cytoplasmic-genic male-sterility (CMS) sterility maintainer inbred lines. R lines are CMS fertility restorer inbred lines.

1999) ranging from 0.14 to 0.89 (mean = 0.50) and the minimum number of recombination events ranging from 0 to 9 (mean = 2.9), nonlinear regression revealed relatively slow LD decay in modern cultivars. LD (quantified by r^2) was still in the neighborhood of 0.30–0.40 at a distance of 5.5 kbp among inbred line alleles. Predictably, recombination estimates increased and LD decreased when wild alleles were included in the analysis (data not shown).

DISCUSSION

Nucleotide diversity in elite and exotic sunflower: Domestication and breeding create population bottlenecks and erode genetic diversity (Buckler *et al.* 2001; Tenaillon *et al.* 2001, 2002; Yamasaki *et al.* 2005; Doebley *et al.* 2006). While genetic diversity has been narrowed by both processes in sunflower (Tang and Knapp 2003; Harter *et al.* 2004; Liu and Burke 2006), diverse and complex parentage and migration (Cheres and Knapp 1998) have apparently partially counteracted the effects of domestication and other diversity-reducing processes in modern oilseed sunflower inbred lines. Significant nucleotide diversity was discovered across inbred lines despite the effects of genetic drift and the winnowing of unfavorable alleles through intense selection and inbreeding in single-cross hybrid sunflower breeding programs. The inbred lines surveyed here retained more than 70% of the nucleotide diversity found in wild progenitors, $\theta = 0.0094$ in elite inbred lines *vs.* $\theta = 0.0128$ in wild progenitors (Table 3) (Liu and Burke 2006). Surprisingly, nucleotide diversity was estimated to be 1.7-fold greater in elite inbred lines than primitive and early open-pollinated (OP) cultivars ($\theta = 0.0056$) (Tables 2 and 3) (Liu and Burke 2006). While the latter estimate was based on data from a smaller number of genes, this finding suggests that the land races and early OP cultivars supplied only a fraction of the genetic diversity found in elite inbred lines.

The germplasm underlying modern oilseed sunflower cultivars was not founded by direct selection in primitive and early OP cultivars alone, but through breeding in elite and exotic germplasm (Cheres and Knapp 1998). Although the early history of sunflower breeding is incomplete, our data support the notion that genetic diversity in modern cultivars has been supplemented by the introgression of wild and exotic

TABLE 3

Nucleotide diversity among 10 sunflower inbred line alleles resequenced from 71 genic loci

Source	Nucleotides sequenced/genotype (kbp)	No. of polymorphic sites (s)	DNA polymorphism frequency (bp/s)	Mean no. of segregating sites (θ)	Mean no. of pairwise dequence differences (π)
Nucleotide polymorphisms	49.4	1256	39.3	0.0094 ± 0.0043	0.0107 ± 0.0058
INDEL polymorphisms	49.4	178	277.0	0.0013 ± 0.0006	0.0016 ± 0.0009
SNPs	49.4	1078	45.7	0.0081 ± 0.0037	0.0092 ± 0.0049
Coding sequences	26.4	420	62.8	0.0059 ± 0.0027	0.0063 ± 0.0034
Synonymous substitutions	6.2	275	22.5	0.0164 ± 0.0075	0.0176 ± 0.0095
Nonsynonymous substitutions	20.2	145	139.4	0.0026 ± 0.0012	0.0028 ± 0.0016
Noncoding sequences	20.2	628	32.1	0.0115 ± 0.0052	0.0135 ± 0.0073
5′ UTR sequences	1.3	30	42.3	0.0087 ± 0.0043	0.0111 ± 0.0062
Intron sequences	12.3	422	29.2	0.0126 ± 0.0058	0.0148 ± 0.0080
3′ UTR sequences	6.6	176	37.5	0.0098 ± 0.0045	0.0116 ± 0.0063

Nucleotide diversity was surveyed in 107 resequenced amplicons (RSAs) amplified from 71 RFLP marker loci genotyped using 67 cDNA and 4 genomic DNA probes. The number of inbred line allele sequences/locus ranged from 6 to 10. Statistics for coding and noncoding sequences were estimated from 103 RSAs amplified from the 67 cDNA-RFLP marker loci.

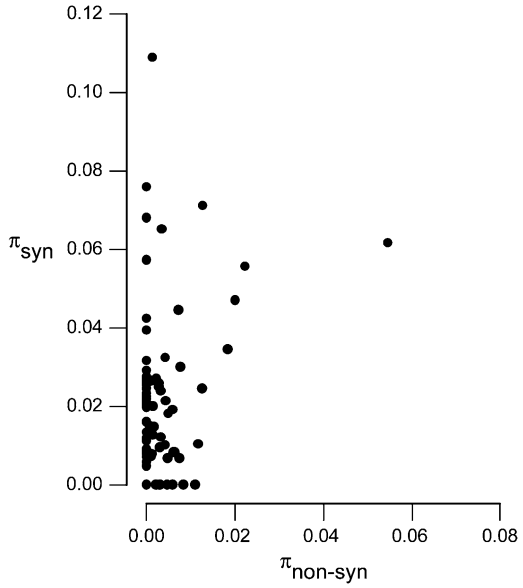

FIGURE 2.—Nucleotide diversities for synonymous (π_{syn}) and nonsynonymous (π_{nonsyn}) SNPs among sunflower 10 inbred line alleles (107 RSAs), resequenced from 71 genic loci distributed among the 17 linkage groups of sunflower ($2n = 2x = 34$).

alleles. Because the sunflower domestication syndrome is complex, the number of loci under selection in wide hybrids in contemporary oilseed sunflower breeding programs is predicted to be large; at least 14 of the 17 chromosomes are known to harbor phenotypic and quantitative trait loci for domestication and confectionery traits and should be under strong selection in oilseed sunflower breeding programs (BURKE *et al.* 2002, 2005; GANDHI *et al.* 2005; TANG *et al.* 2006a). The introgression of wild alleles into modern oilseed sunflower inbred lines has produced a patchwork of elite and wild alleles. Unique haplotypes were found in one or more inbred lines for several of the loci sampled (Figure 5; supplemental Figure 1 at http://www.genetics.org/supplemental/). As noted earlier, sampling may partly underlie differences between the present analysis and the work of LIU and BURKE (2006). Here, we resequenced alleles from a sample of 81 previously mapped genic loci and performed analyses on 107 fragments amplified from 71 loci (BERRY *et al.* 1995; GEDIL *et al.* 2001), whereas LIU and BURKE (2006) resequenced alleles from a random sample of nine genic loci. Moreover, because RFLP markers for the former were known to be polymorphic among oilseed inbred lines (BERRY *et al.* 1994, 1995), the resequenced loci could be more polymorphic, as a whole, than a random sample of loci. As a point of comparison, SSRs revealed greater diversity in land races than oilseed inbred lines (TANG and KNAPP 2003; HARTER *et al.* 2004).

Nucleotide diversity in autogamous and allogamous plant species: Nucleotide diversity in sunflower is slightly lower than maize (REMINGTON *et al.* 2001; TENAILLON *et al.* 2001, 2002; CHING *et al.* 2002; LIU and BURKE 2006; BUCKLER *et al.* 2006), two- to fivefold greater than other domesticated grasses (BUCKLER *et al.* 2001), eight- to 10-fold greater than soybean (ZHU *et al.* 2003; VAN *et al.* 2005), and several-fold greater than other autogamous plant species (KANAZIN *et al.* 2002; GARRIS *et al.* 2003; HAMBLIN *et al.* 2004). Observed SNP frequencies seem to be comparable in sunflower and

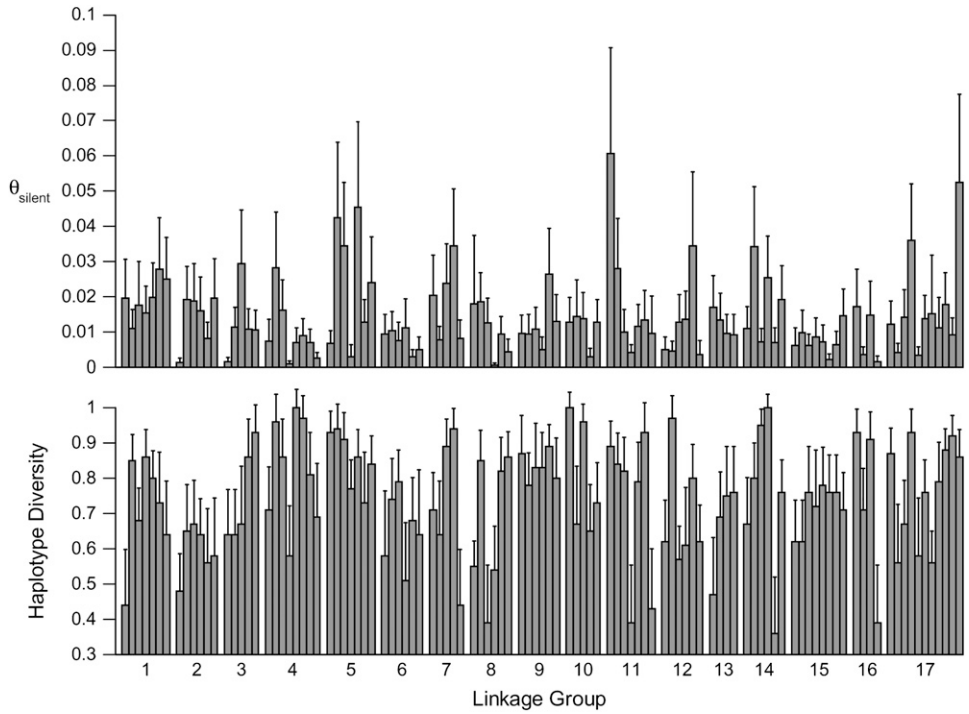

FIGURE 3.—θ_{silent} and haplotype diversity statistics among 10 inbred line and two wild alleles amplified from 71 genic loci (107 RSAs) distributed among the 17 linkage groups of sunflower ($2n = 2x = 34$). Loci are displayed in the order found on each linkage group from 1 (top) to 17 (bottom).

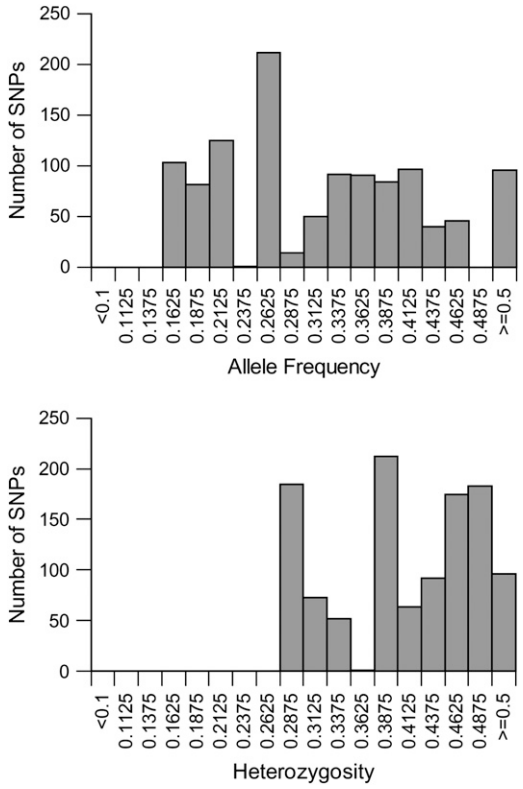

FIGURE 4.—SNP allele frequency (least common allele) and heterozygosity distributions for 1078 SNPs identified in 10 inbred line alleles resequenced from 71 genic loci (107 RSAs).

maize inbred lines. SNP frequencies were 1/32 bp in noncoding and 1/63 bp in coding sequences in sunflower inbred lines (Table 3) and 1/31 bp in noncoding and 1/124 bp in coding sequences in maize inbred lines

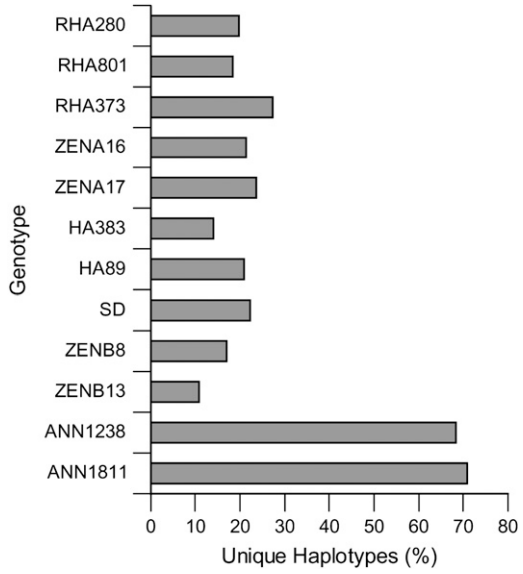

FIGURE 5.—Percentage of unique haplotypes identified among inbred line and wild alleles resequenced from 71 genic loci (107 RSAs).

(CHING *et al.* 2002). SNP frequencies, however, are sensitive to the number of genotypes sampled (larger samples have a greater likelihood of capturing rare SNPs), and the studies referenced above differed widely in terms of sampling strategies. However, because θ is roughly proportional to heterozygosity, the expected number of nucleotide differences between a randomly selected pair of alleles can be estimated. For sunflower, a randomly selected pair of elite alleles is expected to differ at 1 out of every 106 nucleotides (*i.e.*, $1/0.0094 = 106.4$), whereas corn is expected to differ at 1 out of every 105 nucleotides (TENAILLON *et al.* 2001), and soybean is expected to differ at 1 out of every 1,030 nucleotides (ZHU *et al.* 2003). Hence, SNP frequencies seem to be sufficient in the modern sunflower cultivars for the development of SNP genotyping assays for most loci and for very high density genetic mapping using highly parallel SNP genotyping methods (BOREVITZ *et al.* 2003; HAZEN and KAY 2003; WINZELER *et al.* 2003; WERNER *et al.* 2005; GUNDERSON *et al.* 2006; SYVANEN 2001, 2005).

SNP abundance in sunflower and other plant genomes: The genic loci we sampled supply an estimate of the number of common SNPs in the sunflower genome. With 3500 Mbp of DNA in the nuclear genome (BAACK *et al.* 2005) and 1078 SNPs in the 49.4 kbp sample of DNA surveyed in the present study (Table 3), modern sunflower cultivars are predicted to harbor at least 76.4 million SNPs (3,500,000,000 bp/49,400 bp × 1078 SNPs). When translated into genetic distance, modern cultivars are predicted to harbor at least 54,571 SNPs/cM, assuming 1400 cM in the sunflower genome (TANG *et al.* 2002; YU *et al.* 2002, 2003). These estimates assume the loci sampled are typical of DNA as a whole in sunflower and do not account for rare SNPs below the threshold of detection in our study ($f_r < 0.125$). If the inbred lines we selected under represent allelic diversity in modern cultivars, and the protein coding loci we selected are less polymorphic than noncoding DNA in sunflower, the number of common SNPs will be >76.4 million. Conversely, if the loci selected for resequencing are more polymorphic than the balance of the genome, 76.4 million may overestimate the number of common SNPs in modern cultivars. Cultivated soybean, which is significantly less polymorphic than cultivated sunflower, is predicted to harbor 4–5 million SNPs in 1115 Mbp of DNA (ZHU *et al.* 2003; YOON *et al.* 2007), whereas maize inbred lines, with 114 SNPs in 6935 bp of DNA (CHING *et al.* 2002), is predicted to harbor 41 million SNPs in 2500 Mbp of DNA. The number of SNPs in sunflower, per Mbp of DNA (21,800/Mbp), is estimated to be five- to sixfold greater than soybean (3587–4484/Mbp) and 1.3-fold greater than maize (16,400/Mbp). Hence, the predicted number of SNPs in cultivated and wild sunflower is on par with the most polymorphic plant species surveyed so far (BUCKLER *et al.* 2001; REMINGTON *et al.* 2001; TENAILLON *et al.* 2001, 2002; CHING *et al.* 2002;

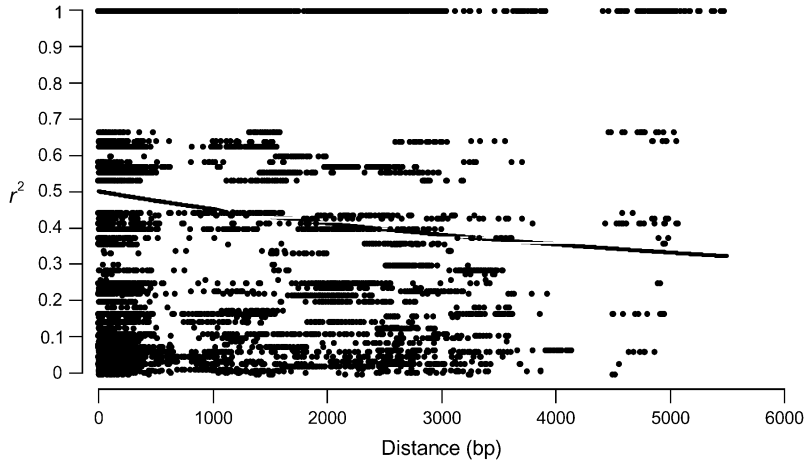

FIGURE 6.—Squared allele frequency correlations (r^2) as a function of physical distance (bp) among polymorphic sites identified in alleles resequenced from 10 inbred lines. The predicted decline in LD (solid line) was found by nonlinear regression of r^2 on bp using the mutation-recombination-drift model of HILL and WEIR (1988). See text for details.

KANAZIN *et al.* 2002; GARRIS *et al.* 2003; HAMBLIN *et al.* 2004; ZHU *et al.* 2003; BUCKLER *et al.* 2006; LIU and BURKE 2006; VAN *et al.* 2005).

Nucleotide and haplotype diversity within and between heterotic groups: Two wild alleles (ANN1238 and ANN1811) were resequenced to supply a benchmark for assessing differences in haplotype structure, SNP frequencies, and nucleotide and haplotype diversities between elite and wild sunflower alleles. Similar to maize (CHING *et al.* 2002), we identified a small number of distinct haplotypes (one to nine) among inbred line alleles, where intralocus SNPs comprising haplotypes were in LD (supplemental Figure 1 at http://www.genetics.org/supplemental/). Selection for seed yield and hybrid seed production traits has created broad female (B) and male (R) heterotic groups in sunflower (BERRY *et al.* 1994; GENTZBITTEL *et al.* 1994; HONGTRAKUL *et al.* 1997; CHERES and KNAPP 1998). Significant genetic diversity has apparently been preserved in a small number of highly divergent B- and R-line haplotypes in sunflower, where haplotype divergence is greater between than within heterotic groups (supplemental Figure 1). While heterotic groups seem to be much less sharply differentiated in sunflower than maize, patterns of genetic diversity and haplotype divergence seem to be similar within and between heterotic groups in both species (TENAILLON *et al.* 2001, 2002; YU *et al.* 2002, 2003; LIU *et al.* 2003; REIF *et al.* 2003; JUNG *et al.* 2004; CHING *et al.* 2002). By contrast, haplotypes seem to be unstructured in the wild progenitor of maize (WHITE and DOEBLEY 1999; LIU *et al.* 2003) and wild sunflower (SLABAUGH *et al.* 2003; TANG and KNAPP 2003; KOLKMAN *et al.* 2004; LIU and BURKE 2006). While we only sampled two wild alleles/locus, wild haplotypes for two-thirds of the loci were unique (Figure 5; supplemental Figure 1).

Heterozygosity and haplotype diversity: SNPs and other biallelic DNA markers are, as a whole, less informative than mulitallelic RFLP and SSR markers; however, when multiple SNPs in haplotype blocks are genotyped, the informativeness of SNP haplotypes should be comparable to RFLP and SSR markers (CHING *et al.* 2002). The inbred lines selected for allele resequencing (Table 2) were predicted from pedigree and RFLP, AFLP, and SSR marker diversity analyses to broadly sample genetic diversity, capture a significant percentage of the nucleotide diversity in elite inbred lines, and to be minimally redundant (BERRY *et al.* 1994; GENTZBITTEL *et al.* 1994; CHERES and KNAPP 1998; GEDIL *et al.* 2001; YU *et al.* 2002, 2003; TANG and KNAPP 2003). SNP heterozygosities and haplotype diversities were therefore expected to be greater among the resequenced inbred line alleles than among a random sample of inbred line alleles (Figures 3 and 4; supplemental Figure 1 at http://www.genetics.org/supplemental/). The probability of observing RFLP or SSR polymorphisms between two inbred lines (h_p) has been in the 0.32–0.53 range in several inbred line surveys in sunflower (BERRY *et al.* 1994; GENTZBITTEL *et al.* 1994; YU *et al.* 2002a,b; TANG and KNAPP 2003). The probability of observing different SNP haplotypes (p_s) between two inbred lines was 0.57 in the present study and thus slightly greater than h_p for RFLP and SSR markers (supplemental Figure 1). The difference could be an artifact of sampling differences; we selected inbred lines to minimize redundancy and maximize uniqueness, whereas several inbred lines within heterotic groups were sampled in previous RFLP and SSR diversity surveys, thereby increasing redundancy and decreasing heterozygosity. With deeper sampling, haplotype diversity should decrease, whereas the number of haplotypes should not substantially increase (CHING *et al.* 2002; ZHU *et al.* 2003; VAN *et al.* 2005); deeper sampling is predicted to identify less common alleles introgressed into elite inbred lines from exotic germplasm sources.

LD: The rate of decay of LD affects the resolution of association mapping analyses and the density of DNA markers needed for identifying phenotype–genotype associations (JORDE 1995, 2000; BUCKLER and THORNSBERRY 2002; NORDBORG *et al.* 2002; CHING *et al.* 2002; RAFALSKI and MORGANTE 2004; BUCKLER *et al.* 2006). The rapid

decay of LD in wild sunflower and maize alleles (Ching *et al.* 2002; Liu and Burke 2006) facilitates very high-resolution association mapping; however, concomitantly high DNA marker densities are needed for discovering associations (Risch 2000; Cardon and Bell 2001; Johnson *et al.* 2001; Stumpf 2002; Greenwood *et al.* 2004; Weigel and Nordborg 2005; Kim *et al.* 2006). Lower DNA marker densities are needed for association mapping in species where LD persists over greater physical distances, although resolution decreases (Cardon and Abecasis 2003). Our results indicate that LD persists over longer tracts of DNA in inbred lines than primitive and early open-pollinated cultivars and wild populations in sunflower (Liu and Burke 2006). LD decayed to $r^2 = 0.1$ by 200 bp in wild populations and 1100 bp in OP cultivars (Liu and Burke 2006), but only decayed to 0.32 by 5500 bp in inbred lines in our study, the longest physical distance surveyed (Figure 6); analyses of longer tracts of DNA are needed to more thoroughly assess LD decay in inbred lines. While there was significant LD variability among loci, the slower decay in sunflower inbred lines can be attributed to population bottlenecks produced by inbreeding and artificial selection, a common phenomenon in domesticated species where intense selection has been practiced for many generations (Buckler *et al.* 2001; Ching *et al.* 2002; Doebley *et al.* 2006). Whether analyses are done in domesticated or wild germplasm, very high DNA marker densities are needed for association mapping in sunflower, a species with ample diversity to support such analyses.

This research was supported by grants from the United States Department of Agriculture National Research Initiative Plant Genome Program (no. 2000-04292), the National Science Foundation Plant Genome Program (no. 0421630), and Advanta Semillas.

LITERATURE CITED

Altschul, S. F., and W. Gish, 1996 Local alignment statistics. Meth. Enzymol. **266:** 460–480.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. **215:** 403–410.

Aquadro C. F., V. Bauer DuMont and F. A. Reed, 2001 Genome-wide variation in the human and fruitfly: a comparison. Curr. Opin. Genet. Dev. **11:** 627–634.

Baack, E. J., K. D. Whitney and L. H. Rieseberg, 2005 Hybridization and genome size evolution: timing and magnitude of nuclear DNA content increases in *Helianthus* homoploid hybrid species. New Phytol. **167:** 623–630.

Barnes, W. M., 1994 PCR amplification of up to 35-kb DNA with high fidelity and high yield from bacteriophage templates. Proc. Natl. Acad. Sci. USA **91:** 2216–2220.

Berry, S. T., R. J. Allen, S. R. Barnes and P. D. S. Caligari, 1994 Molecular marker analysis of *Helianthus annuus* L. 1. Restriction fragment length polymorphisms between inbred lines of cultivated sunflower. Theor. Appl. Genet. **89:** 435–441.

Berry, S. T., A. J. Leon, C. C. Hanfrey, P. Challis, A. Burkholz *et al.*, 1995 Molecular marker analysis of *Helianthus annuus* L. 2. Contruction of an RFLP linkage map for cultivated sunflower. Theor. Appl. Genet. **91:** 195–199.

Bhattramakki, D., M. Dolan, M. Hanafey, R. Wineland, D. Vaske *et al.*, 2002 Insertion-deletion polymorphisms in 3′ regions of maize genes occur frequently and can be used as highly informative genetic markers. Plant Mol. Biol. **48:** 539–547.

Borevitz, J. O., D. Liang, D. Plouffe, H. S. Chang, T. Zhu *et al.*, 2003 Large-scale identification of single-feature polymorphisms in complex genomes. Genome Res. **13:** 513–523.

Buckler, E. S., and J. M. Thornsberry, 2002 Plant molecular diversity and applications to genomics. Curr. Opin. Plant Biol. **5:** 107–111.

Buckler, E. S., J. M. Thornsberry and S. Kresovich, 2001 Molecular diversity, structure and domestication of grasses. Genet. Res. **77:** 213–218.

Buckler, E. S., B. S. Gaut and M. D. McMullen, 2006 Molecular and functional diversity of maize. Curr. Opin. Plant. Biol. **9:** 172–176.

Burke, J. H., S. J. Knapp and L. H. Rieseberg, 2005 Genetic consequences of selection during the evolution of cultivated sunflower. Genetics **171:** 1933–1940.

Burke, J. M., S. Tang, S. J. Knapp and L. H. Rieseberg, 2002 Genetic analysis of sunflower domestication. Genetics **161:** 1257–1267.

Cardon, L. R., and G. R. Abecasis, 2003 Using haplotype blocks to map human complex loci. Trends Genet. **19:** 135–140.

Cardon, L. R., and J. I. Bell, 2001 Association study designs for complex diseases. Nat. Rev. Genet. **2:** 91–99.

Chapman, N. H., and E. A. Thompson, 2001 Linkage disequilibrium mapping: the role of population history, size, and structure. Adv. Genet. **42:** 413–437.

Cheres, M., and S. J. Knapp, 1998 Ancestral origins and genetic diversity of cultivated sunflower: coancestry analysis of public germplasm sources. Crop Sci. **38:** 1476–1482.

Ching, A., K. S. Caldwell, M. Jung, M. Dolan, O. S. Smith *et al.*, 2002 SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet. **3:** 19.

Collins, F. S., L. D. Brooks and A. Chakravarti, 1998 A DNA polymorphism discovery resource for research on human genetic variation. Genome Res. **8:** 1229–1231.

Doebley, J. F., B. S. Gaut and B. D. Smith, 2006 The molecular genetics of crop domestication. Cell **127:** 1309–1321.

Ewing, B., and P. Green, 1998 Basecalling of automated sequencer traces using phred. II. Error probabilities. Genome Res. **8:** 186–194.

Ewing, B., L. Hillier, M. Wendl and P. Green, 1998 Basecalling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. **8:** 175–185.

Flint-Garcia, S. A., J. M. Thornsberry and E. S. Bucker, 2003 Structure of linkage disequilibrium in plants. Annu. Rev. Plant. Biol. **54:** 357–374.

Gandhi, S. D., A. F. Heesacker, C. A. Freeman, J. Argyris, K. Bradford *et al.*, 2005 The self-incompatibility locus (S) and quantitative trait loci for self-pollination and seed dormancy in sunflower. Theor. Appl. Genet. **111:** 619–629.

Garris, A. J., S. R. McCouch and S. Kresovich, 2003 Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the *xa5* locus of rice (*Oryza sativa* L.). Genetics **165:** 759–769.

Gedil, M. A., C. Wye, S. Berry, B. Segers, J. Peleman *et al.*, 2001 An integrated restriction fragment length polymorphism–amplified fragment length polymorphism linkage map for cultivated sunflower. Genome **44:** 213–221.

Gentzbittel, L., Y. X. Zhang, F. Vear, B. Griveau and P. Nicolas, 1994 RFLP studies of genetic relationships among inbred lines of the cultivated sunflower, *Helianthus annuus* L.: evidence fore distinct restorer and maintainer germplasm pools. Theor. Appl. Genet. **92:** 419–425.

Greenwood, T. A., B. K. Rana and N. J. Schork, 2004 Human haplotype block sizes are negatively correlated with recombination rates. Genome Res. **14:** 1358–1361.

Gunderson, K. L., F. J. Steemers, H. Ren, P. Ng, L. Zhou *et al.*, 2006 Whole-genome genotyping. Methods Enzymol. **410:** 359–376.

Halushka, M. K., J. B. Fan, K. Bentley, L. Hsie, N. Shen *et al.*, 1999 Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat. Genet. **22:** 239–247.

Hamblin, M. T., S. E. Mitchell, G. M. White, J. Gallego, R. Kukatla *et al.*, 2004 Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of *Sorghum bicolor.* Genetics **167:** 471–483.

Harter, A. V., K. A. Gardner, D. Falush, D. L. Lentz, R. A. Bye *et al.*, 2004 Origin of extant domesticated sunflowers in eastern North America. Nature **430:** 201–205.

Hass, C., S. Tang, S. Leonard, J. F. Miller, M. Traber *et al.*, 2006 Three non-allelic epistatically interacting methyltransferase mutations produce novel tocopherol (vitamin E) profiles in sunflower. Theor. Appl. Genet. **113:** 767–782.

Hazen, S. P., and S. A. Kay, 2003 Gene arrays are not just for measuring gene expression. Trends Plant Sci. **8:** 413–416.

Hill, W. G., and B. S. Weir, 1988 Variances and covariances of squared linkage disequilibria in finite populations. Theor. Popul. Biol. **33:** 54–78.

Hongtrakul, V., G. Huestis and S. J. Knapp, 1997 Amplified fragment length polymorphisms as a tool for DNA fingerprinting sunflower germplasm: genetic diversity among oilseed inbred lines. Theor. Appl. Genet. **95:** 400–407.

Hudson, R. R., 2001 Linkage disequilibrium and recombination, pp. 309–324 in *Handbook of Statistical Genetics*, edited by D. J. Balding, M. Bishop and C. Cannings. John Wiley and Sons, Chichester, UK.

Hudson, R. R., and N. L. Kaplan, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

Ingvarsson, P. K., 2005 Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., Salicaceae). Genetics **169:** 945–953.

Johnson, G. C., L. Esposito, B. J. Barratt, A. N. Smith, J. Heward *et al.*, 2001 Haplotype tagging for the identification of common disease genes. Nat. Genet. **29:** 233–237.

Jorde, L. B., 1995 Linkage diseqilibrium as a gene-mapping tool. Am. J. Hum. Genet. **56:** 11–14.

Jorde, L. B., 2000 Linkage disequilibrium and the search for complex disease genes. Genome Res. **10:** 1435–1444.

Jung, M., A. Ching, D. Bhattramakki, M. Dolan, S. Tingey *et al.*, 2004 Linkage disequilibrium and sequence diversity in a 500-kbp region around the *adh1* locus in elite maize germplasm. Theor. Appl. Genet. **109:** 681–689.

Kanazin, V., H. Talbert, D. See, P. DeCamp, E. Nevo *et al.*, 2002 Discovery and assay of single-nucleotide polymorphisms in barley (*Hordeum vulgare*). Plant. Mol. Biol. **48:** 529–537.

Kim, S., K. Zhao, R. Jiang, J. Molitor, J. O. Borevitz *et al.*, 2006 Association mapping with single-feature polymorphisms. Genetics **173:** 1125–1133.

Kolkman, J. M., M. B. Slabaugh, J. M. Bruniard, S. T. Berry, S. B. Bushman *et al.*, 2004 Acetohydroxyacid synthase mutations conferring resistance to imidazolinone or sulfonylurea herbicides in wild sunflower biotypes. Theor. Appl. Genet. **109:** 1147–1159.

Lindblad-Toh K., E. Winchester, M. Daly, D. Wang, J. N. Hirschhorn *et al.*, 2000 Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. Nat. Genet. **24:** 381–386.

Liu, A., and J. M. Burke, 2006 Patterns of nucleotide diversity in wild and cultivated sunflower. Genetics **173:** 321–330.

Liu, K., M. Goodman, S. Muse, J. S. Smith, E. Buckler *et al.*, 2003 Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics **165:** 2117–2128.

McGinnis, S., and T. L. Madden, 2004 BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. **32:** W20–W25.

Murray, M. G., and W. R. Thompson, 1980 Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res. **8:** 4321–4325.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nordborg, M., and S. Tavare, 2002 Linkage disequilibrium: what history has to tell us. Trends Genet. **18:** 83–90.

Nordborg, M., J. O. Borevitz, J. Bergelson, C. C. Berry, J. Chory *et al.*, 2002 The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nature Genet. **30:** 190–193.

Rafalski, A, 2002a Applications of single nucleotide polymorphisms in crop genetics. Curr. Opinion Plant Biol. **5:** 94–100.

Rafalski, A., 2002b Novel genetic mapping tools in plants: SNPs and LD-based approaches. Plant Sci. **162:** 329–333.

Rafalski, A., and M. Morgante, 2004 Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends Genet. **20:** 103–111.

Reif, J. C., A. E. Melchinger, X. C. Xia, M. L. Warburton, D. A. Hoisington *et al.*, 2003 Use of SSRs for establishing heterotic groups in subtropical maize. Theor. Appl. Genet. **107:** 947–957.

Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt *et al.*, 2001 Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA **98:** 11479–11484.

Risch, N. J., 2000 Searching for genetic determinants for the new millenium. Nature **405:** 847–856.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

Rozas, J., J. C. Sánchez-DelBarrio, X. Messegyer and R. Rozas, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:** 2496–2497.

Schuppert, G. F., S. Tang, M. B. Slabaugh and S. J. Knapp, 2006 The sunflower high-oleic mutant *Ol* carries variable tandem repeats of *FAD2–1*, a seed-specific oleoyl-phosphatidyl choline desaturase. Mol. Breeding **17:** 241–256.

Shifman, S., J. Kuypers, M. Kokoris, B. Yakir and A. Darvasi, 2003 Linkage disequilibrium patterns of the human genome across populations. Hum. Mol. Genet. **12:** 771–776.

Slabaugh, M. B., J. K. Yu, S. Tang, A. Heesacker, X. Hu *et al.*, 2003 Haplotyping and mapping a large cluster of downy mildew resistance gene candidates in sunflower using multilocus intron fragment length polymorphisms. Plant Biotechnol. J. **1:** 167–185.

Stumpf, M. P., 2002 Haplotype diversity and the block structure of linkage disequilibrium. Trends Genet. **18:** 226–228.

Syvanen, A. C., 2001 Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat. Rev. Genet. **2:** 930–942.

Syvanen, A. C., 2005 Toward genome-wide SNP genotyping. Nat. Genet. **37:** S5–S10.

Tang, S., and S. J. Knapp, 2003 Microsatellites uncover extraordinary diversity in native American landraces and wild populations of cultivated sunflower. Theor. Appl. Genet. **106:** 990–1003.

Tang, S., J.-K. Yu, M. B. Slabaugh, D. K. Shintani and S. J. Knapp, 2002 Simple sequence repeat map of the sunflower genome. Theor. Appl. Genet. **105:** 1124–1136.

Tang, S., A. Leon, W. C. Bridges and S. J. Knapp, 2006a Quantitative trait loci for genetically correlated seed traits are tightly linked to branching and pericarp pigment loci in sunflower. Crop Sci. **46:** 721–734.

Tang, S., C. Hass and S. J. Knapp, 2006b *Ty3/gypsy*-like retrotransposon knockout of a 2-methyl-6-phytyl-1,4-benzoquinone methyltransferase is non-lethal, unmasks a cryptic paralogous mutation, and produces novel tocopherol (vitamin E) profiles in sunflower. Theor. Appl. Genet. **113:** 783–799.

Taramino, G., and S. Tingey, 1996 Simple sequence repeats for germplasm analysis and mapping in maize. Genome **39:** 277–287.

Tenaillon, M., M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley *et al.*, 2001 Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc. Natl. Acad. Sci. USA **98:** 9161–9166.

Tenaillon, M. I., M. C. Sawkins, L. K. Anderson, S. M. Stack, J. Doebley *et al.*, 2002 Patterns of diversity and recombination along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Genetics **162:** 1401–1413.

Van, K., E. Y. Hwang, M. Y. Kim, H. J. Park, S. H. Lee *et al.*, 2005 Discovery of SNPs in soybean genotypes frequently used as the parents of mapping populations in the United States and Korea. J. Hered. **96:** 529–535.

Wall, J. D., 1999 Recombination and the power of statistical tests of neutrality. Genet. Res. **74:** 65–79.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Weigel, D., and M. Nordborg, 2005 Natural variation in Arabidopsis. How do we find the causal genes? Plant Physiol. **138:** 567–568.

Weir, B. S., 1996 *Genetic Data Analysis II*. Sinauer, Sunderland, MA.

Werner, J. D., J. O. Borevitz, N. H. Uhlenhaut, J. R. Ecker, J. Chory *et al.*, 2005 FRIGIDA-independent variation in flowering time of natural *A. thaliana* accessions. Genetics **170:** 1197–1207.

White, S. E, and J. F. Doebley, 1999 The molecular evolution of *terminal ear 1*, a regulatory gene in the genus *Zea*. Genetics **153:** 1455–1462.

Wiltshire, R., M. T. Pletcher, S. Batalov, S. W. Barnes, L. M. Tarantino *et al.*, 2003 Genome-wide single-nucleotide polymorphism analysis defines haplotype patterns in mouse. Proc. Natl. Acad. Sci. USA **100:** 3380–3385.

Winzeler, E. A., C. I. Castillo-Davis, G. Oshiro, D. Liang, D. R. Richards *et al.*, 2003 Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. Genetics **163:** 79–89.

Yamasaki, M., M. I. Tenaillon, S. G. Schroeder, H. Sanchez-Villeda, J. F. Doebley *et al.*, 2005 A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. Plant Cell **17:** 2859–2872.

Yoon M. S., Q. J. Song, I. Y. Choi, J. E. Specht, D. L. Hyten *et al.*, 2007 BARCSoySNP23: a panel of 23 selected SNPs for soybean cultivar identification. Theor. Appl. Genet. **114:** 885–899.

Yu, J. K., J. Mangor, L. Thompson, K. J. Edwards, M. B. Slabaugh *et al.*, 2002 Allelic diversity of simple sequence repeat markers among elite inbred lines in cultivated sunflower. Genome **45:** 652–660.

Yu, J. K., S. Tang, M. B. Slabaugh, A. Heesacker, G. Cole *et al.*, 2003 Towards a saturated molecular genetic linkage map for cultivated sunflower. Crop Sci. **43:** 367–387.

Zhu, Y. L., Q. J. Song, D. L. Hyten, C. P. Van Tassell, L. K. Matukumalli *et al.*, 2003 Single-nucleotide polymorphisms in soybean. Genetics **163:** 1123–1134.

Communicating editor: B. J. Walsh