Theor Appl Genet (2008) 117:1021–1029
DOI 10.1007/s00122-008-0841-0

ORIGINAL PAPER

SSRs and INDELs mined from the sunflower EST database: abundance, polymorphisms, and cross-taxa utility

Adam Heesacker · Venkata K. Kishore · Wenxiang Gao · Shunxue Tang · Judith M. Kolkman · Alan Gingle · Marta Matvienko · Alexander Kozik · Richard M. Michelmore · Zhao Lai · Loren H. Rieseberg · Steven J. Knapp

Received: 14 June 2007 / Accepted: 28 June 2008 / Published online: 17 July 2008

Abstract Simple sequence repeats (SSRs) are abundant and frequently highly polymorphic in transcribed sequences and widely targeted for marker development in eukaryotes. Sunflower (*Helianthus annuus*) transcript assemblies were built and mined to identify SSRs and insertions-deletions (INDELs) for marker development, comparative mapping, and other genomics applications in sunflower. We describe the spectrum and frequency of SSRs identified in the sunflower EST database, a catalog of 16,643 EST-SSRs, a collection of 484 EST-SSR and 43 EST-INDEL markers developed from common sunflower ESTs, polymorphisms of the markers among the parents of several intraspecific and interspecific mapping populations, and the transferability of the markers to closely and distantly related species in the Compositae. Of 17,904 unigenes in the transcript assembly, 1,956 (10.9%) harbored one or more SSRs with repeat counts of $n \geq 5$. EST-SSR markers were 1.6-fold more polymorphic among exotic than elite genotypes and 0.7-fold less polymorphic than non-genic SSR markers. Of 466 EST-SSR or INDEL markers screened for cross-species amplification and polymorphisms, 413 (88.6%) amplified alleles from one or more wild species (*H. argophyllus*, *H. tuberosus*, *H. anomalus*, *H. paradoxus*, and *H. deserticola*), whereas 69 (14.8%) amplified alleles from safflower (*Carthamus tinctorius*) and 67 (14.4%) amplified alleles from lettuce (*Lactuca sativa*); hence, only a fraction were transferable to distantly related genera in the Compositae, whereas most were transferable to wild relatives of *H. annuus*. Several thousand additional SSRs were identified in the EST database and supply a wealth of templates for EST-SSR marker development in sunflower.

Communicated by M. Sorrells.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-008-0841-0) contains supplementary material, which is available to authorized users.

A. Heesacker · W. Gao · S. Tang · A. Gingle · S. J. Knapp (✉)
Center for Applied Genetic Technologies,
The University of Georgia, 111 Riverbend Road,
Athens, GA 30602, USA
e-mail: sjknapp@uga.edu

V. K. Kishore · J. M. Kolkman
Department of Crop and Soil Science,
Oregon State University, Corvallis, OR 97331, USA

M. Matvienko · A. Kozik · R. M. Michelmore
Department of Plant Science and Genome Center,
University of California, Davis, CA 95616, USA

Z. Lai · L. H. Rieseberg
Department of Biology, Indiana University,
Bloomington, IN 47405, USA

L. H. Rieseberg
Department of Botany, The University of British Columbia,
Vancouver, BC V6T 1Z4, Canada

Introduction

Simple sequence repeats (SSRs) are abundant and ubiquitious in eukaryotic genomes, frequently multiallelic and highly polymorphic, and, consequently, have been widely targeted for marker development for genetic mapping and other genomic applications in numerous species (Taramino and Tingey 1996; Morgante et al. 2002; Thiel et al. 2003; Yu et al. 2004a, b; Frary et al. 2005; Park et al. 2005; Varshney et al. 2005a, b; Han et al. 2006; Guo et al. 2006; Poncet et al. 2006). SSRs are more frequent in transcribed than non-transcribed sequences and equally frequent in the

transcriptomes of plants with dramatically different nuclear DNA contents (Morgante et al. 2002). Hence, mining EST databases is one of the most expedient approaches for identifying sequences harboring SSRs for the development of highly polymorphic DNA markers. EST databases have been developed for more than 140 biologically and economically important plant species (http://www.ncbi.nih.gov; http://www.tigr.org), including sunflower (*Helianthus annuus* L.), lettuce (*Lactuca sativa* L.), and safflower (*Carthamus tinctorius* L.) (Kozik et al. 2002; http://cgpdb.ucdavis.edu/cgpdb2/; Fernandez et al. 2003; Tamborindeguy et al. 2004; Ben et al. 2005; Lai et al. 2005a). The present study focused on mining the sunflower EST database for SSRs, in addition to assessing their transferability to lettuce and safflower, as a benchmark for assessing cross-amplification among divergent taxa in the Compositae (Asteraceae). EST-SSR markers generally display broad utility within and limited utility among genera in plants, primarily because polymorphisms in sequences flanking repeats increase as phylogenetic distances increase (Peakall et al. 1998; Eujayl et al. 2004; Saha et al. 2004; Guo et al. 2006).

Sunflower (Asteroideae), safflower (Carduoideae), and lettuce (Cichorioideae) are members of different subfamilies in the Compositae (Asteraceae), a cosmopolitan family of 1,600–1,700 genera, 24,000–30,000 species, and numerous agronomically, horticulturally, and medically important species (Jansen et al. 1991; Funk et al. 2005). More than 800,000 ESTs have been developed for *Helianthus*, *Lactuca*, *Carthamus*, and other genera in the family (http://www.ncbi.nlm.nih.gov/), primarily by the Compositae Genome Program (CGP; http://cgpdb.ucdavis.edu/). Before the initial release of ESTs by the CGP, GenBank and other public databases held fewer than 100 sunflower nucleotide sequences. Since then, the CGP has produced 261,699 sunflower ESTs, 284,745 sunflower ESTs have been deposited in GenBank, and numerous sunflower transcript assemblies (TAs) have been built and mined for SSRs and SNPs (Kozik et al. 2002; Gandhi et al. 2005; Lai et al. 2005a; Pashley et al. 2006). The analyses described here focused on the initial collection of 67,180 *H. annuus*, *H. argophyllus*, and *H. paradoxus* ESTs produced by the CGP (GenBank Acc. No. BQ909263-BQ917261, BQ965129-BQ98004, BU015365-BU036497, and CF076145-CF099271) and 22,045 additional *H. annuus* ESTs deposited in GenBank (Fernandez et al. 2003; Tamborindeguy et al. 2004; Ben et al. 2005).

Significant SSR marker resources have been developed for sunflower (Paniego et al. 2002; Tang et al. 2002, 2003; Yu et al. 2002, 2003; Gandhi et al. 2005; Pashley et al. 2006); however, a limited number of SSR and other highly portable DNA markers have been developed for genotyping transcribed loci; thus far, less than 60 EST-SSR markers have been described for sunflower (Gandhi et al. 2005; Pashley et al. 2006). We mined the sunflower EST database for SSRs and insertions-deletions (INDELs) and developed a catalog of EST-SSRs and a collection EST-SSR and INDEL markers for comparative mapping and other genomics applications in sunflower. The abundance and characteristics of SSRs identified in the EST database, polymorphisms of the markers among the parents of several intraspecific and interspecific mapping populations, and the transferability of the markers to wild relatives of *H. annuus*, lettuce, prickly lettuce (*L. serriola* L.), and safflower are described herein.

Materials and methods

Plant materials and DNA isolation

ESTs were produced from two *H. annuus* inbred lines (RHA280 and RHA801), two *H. argophyllus* (silverleaf sunflower; ARG) populations (ARG1834 = PI 494582 and ARG1805 = PI 494571), and one *H. paradoxus* (salt-marsh sunflower; PAR) population (PAR-Cibola). Sixteen *Helianthus*, *Lactuca*, and *Carthamus* germplasm accessions were screened for EST-SSR and INDEL marker amplification and polymorphisms (Table 1). Seeds of RHA280, RHA801, Havasupai, Hopi, ANN1811, TUB-2329, DES-2345, ARG-1834, ARG1805, ANO-2346, Saffire, and Salinas were supplied by the United States Department of Agriculture (USDA) Agricultural Research Service (ARS) National Plant Germplasm System (http://www.ars-grin.gov/npgs/). Seeds of RHA373, RHA377, and HA383 were supplied by the USDA-ARS Northern Crop Science Research Laboratory (Fargo, ND). Seeds of salt-marsh sunflower (PAR-Cibola) and prickly lettuce (92G489) were collected from the wild. Genomic DNA was isolated from leaves harvested from 4 to 6 weeks-old greenhouse grown plants using a modified CTAB (cetyltrimethylammonium bromide) method (Murray and Thompson 1980).

RNA isolation and cDNA library construction

We constructed 26 cDNA libraries, 11 each from RHA280 and RHA801 (common sunflower) and two each from silverleaf and prairie sunflower. Common sunflower RNAs were isolated from callus, roots, shoots, leaves, pre-fertilized flowers, disk and ray flowers, developing kernels, developing hulls (pericarps), chemically induced leaves and roots, germinating seeds, and drought- and heat-stressed leaves, roots, and flowers, silverleaf sunflower RNAs were isolated from non-stressed and drought-stressed seedlings, leaves, and roots, and salt-marsh sunflower RNAs were isolated

Table 1 *Helianthus*, *Carthamus*, and *Lactuca* germplasm accessions screened for EST-SSR and INDEL marker amplification and length polymorphisms

Species	Common name	Plant introduction number	Name	Germplasm group
H. annuus	Common sunflower	PI 552943	RHA280	Inbred line
		PI 599768	RHA801	Inbred line
		PI 560141	RHA373	Inbred line
		PI 560145	RHA377	Inbred line
		PI 578872	HA383	Inbred line
		PI 369359	Hopi	Land race
		PI 369358	Havasupai	Land race
		PI 494567	ANN1811	Wild
H. tuberosus	Jerusalem artichoke	Ames 22229	TUB-2329	Wild
H. deserticola	Desert sunflower	Ames 26094	DES-2345	Wild
H. argophyllus	Silverleaf sunflower	PI 494582	ARG-1834	Wild
H. anomalus	Western sunflower	Ames26095	ANO-2346	Wild
H. paradoxus	Salt Marsh sunflower	–	PAR-Cibola	Wild
C. tinctorius	Safflower	PI 572475	Saffire	Cultivar
L. sativa	Lettuce	PI 536851	Salinas	Cultivar
L. serriola	Prickly lettuce	–	92G489	Wild

from salt-stressed seedlings, leaves, roots, and flowers, as described by Kozik et al. (2002; http://cgpdb.ucdavis.edu/cgpdb2/). RNAs were isolated from tissues ground in liquid nitrogen and resuspended in 1:1 mixture of RNA extraction buffer [0.1 M Tris-HCl, pH 9.0, 0.1 M LiCl, 10 mM EDTA, and 1% (w/v) SDS] and saturated phenol at 70°C. Subsequent to vortexing and centrifugation (15 min, 6,000*g*), the water phase was collected and RNAs were isolated as described by Pawlowski et al. (1994).

Common sunflower and non-stressed silverleaf and salt-marsh sunflower cDNA libraries were constructed using a SMART™ cloning technology (BD Clontech, Palo Alto, CA). Drought-induced and -repressed silverleaf sunflower and salt-induced and -repressed prairie sunflower cDNA libraries were produced by subtraction using PCR-Select Subtraction technology (BD Clontech, Palo Alto, CA), whereby differentially expressed transcripts are enriched by suppression PCR (Diatchenko et al. 1996). For the stress-induced cDNA libraries, the abiotic stress cDNA tester was enriched for differentially expressed transcripts using non-stress cDNA as the driver and vice versa for the stress-repressed cDNA libraries (Kozik et al. 2002). Subsequent to PCR amplification, subtracted cDNAs were cloned into the pGEM vector using the Promega TA cloning method (Madison, WI) and recombinant clones were identified on X-Gal-containing plates.

H. annuus, *H. paradoxus*, and *H. argophyllus* ESTs were produced by Sanger sequencing from the 26 cDNA libraries. The ESTs were processed, trimmed, annotated, and assembled using CGPdb bioinformatic pipelines (http://cgpdb.ucdavis.edu/cgpdb2/). Contigs were screened for SNPs and INDELs using the CGPdb pipeline and unigenes were screened for SSRs using a modified SSR-IT script (http://cgpdb.ucdavis.edu/cgpdb2/; Temnykh et al. 2001). Unigenes in CGPdb transcript assemblies were mined for SSRs and INDELs and used as templates (reference allele sequences) for designing EST-SSR and INDEL marker primers.

EST-SSR and INDEL discovery, marker development, and polymorphism screening

The assembly of 67,180 *H. annuus*, *H. paradoxus*, and *H. argophyllus* ESTs was screened for all possible dinucleotide, trinucleotide, and tetranucleotide repeat motifs using a custom script developed from SSR-IT (Temnykh et al. 2001) with a repeat count (n) threshold of $n \geq 5$. EST contigs were screened for INDELs between RHA280 and RHA801 alleles using the CGPdb Contig Viewer (Kozik et al. 2002). Flanking oligonucleotide primers were designed for 484 SSRs and 43 INDELs using primer 3 (http://www.broad.mit.edu/genome_software/) with manual selection. To facilitate multiplex genotyping on an ABI Prism 3100 Automated Capillary DNA Sequencer (Applied Biosystems, Foster City, CA), forward oligonucleotide primers were labeled with 6FAM or HEX fluorophores (MWG- Biotech, High Point, NC) and target amplicon (reference allele) lengths were chosen to create a uniform distribution of allele lengths in the 100–500 bp range. The 527 primer pairs were screened for amplification and length polymorphisms among 16 germplasm accessions (Table 1) on an ABI3100 using methods described by

Tang et al. (2003). SSR and INDEL allele lengths were scored using GeneMapper (Applied Biosystems, Foster City, CA). The probability of observing a polymorphism between two germplasm accessions drawn at random (heterozygosity = H) was estimated as described by Ott (1999). The effects of SSR type and length on heterozygosity were estimated using SAS PROC MIXED (http://www.sas.com; Cary, NC) with SSR motif type or SSR length as independent variables. Type III F-statistics were tested using orthogonal linear contrasts (CONTRAST statements in PROC MIXED).

Mining the sunflower EST database for SSRs

The abundance and characteristics of SSRs in the sunflower EST database were further assessed by producing and mining an assembly of 89,225 *H. annuus*, *H. argophyllus*, and *H. paradoxus* ESTs downloaded from GenBank dbEST (http://www.ncbi.nlm.nih.gov/dbEST/) on 11-7-2005 (Kozik et al. 2002; Fernandez et al. 2003; Tamborindeguy et al. 2004; Ben et al. 2005). Sunflower ESTs from dbEST were assembled using CAP3 (Huang and Madan 1999). The assembly was processed using custom python scripts, tab-delimited files were generated and deposited in an mySQL database, and BLASTX analyses (Altschul et al. 1990; Altschul and Gish 1996; McGinnis and Madden 2004) were performed against the NCBI Protein Database (http://www.ncbi.nlm.nih.gov/) to infer putative functions of unigenes. Unigenes were screened for the presence of at least one di-, tri-, or tetra-nucleotide repeat using a custom script developed from SSR-IT (Temnykh et al. 2001) with a repeat count threshold of four ($n \geq 4$). *E. coli*, mitochondrial, and vector contaminants were removed from SSR-containing ESTs with Cross_Match (Green 1996). EST and SSR parameters (e.g., motifs, repeats, and species) were uploaded into an Oracle relational database (RDBMS) for subsequent query based analyses. SSR-containing sequences were partitioned by species and separately clustered with PHRAP (Green 1996) to eliminate redundant ESTs. Clustering results and the NCBI-generated unigene collection were uploaded into the RDBMS. mySQL queries were used to produce SSR motif and repeat number distributions.

Results

Mining the sunflower EST database for SSRs and INDELs

We developed and mined an assembly of 44,053 common sunflower, 12,787 silverleaf sunflower (*H. argophyllus* Torr. and A. Gray), and 10,340 salt-marsh sunflower (*H. paradoxus* Heiser) ESTs for SSRs and INDELs (Kozik et al. 2002; GenBank Acc. No. BQ909263- BQ917261, BQ965129- BQ98004, BU015365-BU036497, and CF076145-CF099271). To facilitate the discovery and genetic mapping of SNPs, INDELs, and SSRs in the RHA280 × RHA801 recombinant inbred line (RIL) mapping population (Tang et al. 2002; Yu et al. 2003; Lai et al. 2005a), ESTs were produced from the parents (22,920 from RHA801 and 21,133 from RHA280). The *H. annuus* ESTs assembled into 7,645 singletons and 4,430 contigs (12,075 unigenes), whereas the *H. annuus*, *H. argophyllus*, and *H. paradoxus* ESTs assembled into 11,271 singletons and 6,760 contigs (18,031 unigenes). Putative unigene functions were inferred by BLASTX analyses against the NCBI Protein Database (http://www.ncbi.nlm.nih.gov/) and best hits were compiled in the CGPdb (http://cgpdb.ucdavis.edu/cgpdb2/). The CGPdb Contig Viewer displays DNA polymorphisms in contigs harboring multiple allele sequences and associates ESTs with BLAST annotation and other information. Of 18,031 unigenes in the transcript assembly, 8,295 (46%) had no significant hits, 6,671 (37%) had significant hits to known function proteins, and 3,065 (17%) had significant hits to proteins with no known function. The number of unigenes and percentage of unigenes with known functions were typical of other large-scale EST databases (http://www.ncbi.nlm.nih.gov/dbEST). DNA sequence alignments (contigs), individual ESTs (singletons), BLAST results, and other EST data were deposited and BLAST and keyword search tools were developed for screening EST assemblies in the CGPdb. cDNA clones for the ESTs were deposited at the Arizona Genomics Institute (AGI; http://www.genome.arizona.edu/) for long-term storage and distribution using CGPdb identifiers. Using a repeat count threshold of $n \geq 5$, 2,501 SSRs and 101 INDELs were identified in the *H. annuus-H. argophyllus-H. paradoxus* transcript assembly (18,031 unigenes) and supplied templates for development of the EST-SSR and INDEL marker described herein.

Subsequent to our initial analyses, 22,045 additional *H. annuus* ESTs were deposited in GenBank (Fernandez et al. 2003; Tamborindeguy et al. 2004; Ben et al. 2005). The latter were downloaded and assembled with the *H. annuus*, *H. argophyllus*, and *H. paradoxus* ESTs and the transcript assembly was screened for all possible dinucleotide, trinucleotide, and tetranucleotide repeats using SSR-IT (Temnykh et al. 2001). The 89,225 ESTs assembled into 6,098 contigs and 11,806 singletons (17,904 unigenes). Using a repeat count threshold of $n \geq 4$, 9,854 dinucleotide, 6,189 trinucleotide, and 600 tetranucleotide repeats (16,643 SSRs) were identified and ranged in length from $k = 8$–58 bp (Supplemental Table 1). Using a repeat

count threshold of $n \geq 5$, 2,406 dinucleotide, 2,181 trinucleotide, and 168 tetranucleotide repeats (4,755 SSRs) were identified and ranged in length from $k = 10$–58 bp. The lower repeat count threshold ($n \geq 4$) was used to mine the EST database for SSRs because $n = 4$ trinucleotide and tetranucleotide repeats ($k \geq 12$ in the reference allele sequence) are often polymorphic in sunflower (Yu et al. 2002; Tang and Knapp 2003; Tang et al. 2003). Of the 16,643 $n \geq 4$ SSRs identified, 7,922 were 12 bp or longer (47.6%), 3,354 were 14 bp or longer (20.2%), and 1,253 were 18 bp or longer (7.5%) (Supplemental Table 1). Of the 484 SSRs targeted for marker development in the present study, only eight had repeat counts (n) of four or less. EST-SSR markers were developed for two $n = 3$ tetranucleotide and six $n = 4$ trinucleotide or tetranucleotide repeats (HT287, 293, 474, 512, 536, 998, 1001, and 1002). These SSRs ranged in length from 12 to 16 bp (in reference allele sequences) and were as polymorphic as the $n \geq 5$ EST- SSR markers we developed (minimum, mean, and maximum heterozygosities for the former were 0.12, 0.62, and 0.87, respectively). GenBank accession numbers, unigene identifiers, SSR motifs, repeat counts, and lengths, and other pertinent data for SSRs identified in ESTs are cataloged in Supplemental Table 1 and supply the information needed for developing additional EST-SSR markers for sunflower.

EST-SSR and INDEL marker development, polymorphisms, and cross-taxa utility

We designed primers for 43 INDELs and 484 SSRs identified in *H. annuus* contigs or singletons (Supplemental Tables 2, 3). Forward and reverse primer sequences, allele lengths, SSR repeat motifs and lengths, and other pertinent data for the EST-SSR and INDEL markers (numbered HT276 to HT1058) are shown in Supplemental Table 3. Thirty-nine EST-INDEL and 427 EST-SSR markers produced high quality genoytpes and were screened for length polymorphisms among *Helianthus*, *Carthamus*, and *Lactuca* germplasm accessions (Table 1; allele lengths are shown in Supplemental Table 4). The sunflower germplasm accessions selected for screening are the parents of intraspecific and interspecific mapping populations, have been screened for polymorphisms using SSR markers developed from SSR-enriched genomic DNA libraries (Yu et al. 2002; Tang et al. 2002, 2003), and were selected to assess the transferability of the EST-SSR markers to wild relatives of *H. annuus*.

Null allele frequencies (f_N) were significantly different ($P < 0.0001$) among species (Fig. 1; Supplemental Table 4). The percentage of amplification failures (frequency of null alleles) followed a predictable pattern, with a minimum in common sunflower and maximum in prickly lettuce. The number of amplification failures increased as genetic distance increased and were greatest for safflower, lettuce, and prickly lettuce. f_N ranged from 0.112 in RHA280 (one of the reference allele sources) to 0.258 in ANN1811 (a wild population) among *H. annuus* genotypes and from 0.260 in *H. argophyllus* to 0.345 in *H. tuberosus* among wild sunflower species. The amplification failure percentages for EST-SSR and INDEL markers were nearly identical to those previously reported for non-genic SSR markers among common sunflower germplasm accessions (Tang et al. 2002, 2003). Of the 466 markers, 100% amplified alleles from one or more common sunflower and 88.6% (413) amplified alleles from one or more wild sunflower species, whereas only 13.1–14.8% (61–69) amplified alleles from safflower, lettuce, or prickly lettuce. Collectively, 17.0% (81) amplified alleles from lettuce, prickly lettuce, or both (Fig. 1; Supplemental Table 4).

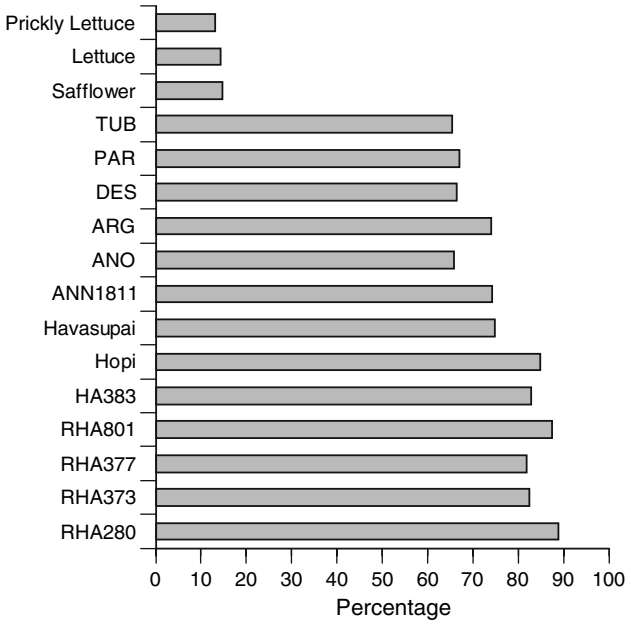

Fig. 1 Cross-taxa amplification percentages for 466 common sunflower EST-SSR and INDEL markers among common and wild sunflower, safflower, lettuce, and prickly lettuce germplasm accessions

EST-SSR and INDEL markers were significantly less polymorphic among elite inbred lines than among exotic and wild germplasm accessions ($P < 0.0001$; Table 2). Heterozygosities for the 466 EST-SSR and INDEL markers ranged from 0.00 to 0.75 among elite oilseed inbred lines, 0.00–0.88 among elite and exotic *H. annuus* germplasm accessions, and 0.00–0.95 among germplasm accessions of common and wild sunflower species (Fig. 2; Supplemental Table 4). The number of monomorphic markers was 2.6-fold greater among elite oilseed inbred lines (209/466) than elite and exotic germplasm accessions of common sunflower (78/466) (Fig. 2; Supplemental Table 4). SSRs and

Table 2 Heterozygosity means for 427 EST-SSR and 39 EST-INDEL markers spanning exon, intron, or UTR sequences, or combinations thereof, among elite, exotic, and wild sunflower germplasm accessions

Source	*H. annuus* elite	*H. annuus* elite + exotic	Elite + exotic + wild[a]
3′ UTR	0.293	0.468	0.634
3′ UTR + intron	0.358	0.506	0.647
5′ UTR	0.283	0.447	0.565
5′ UTR + intron	0.260	0.394	0.625
Exon	0.245	0.417	0.594
Exon + intron	0.321	0.467	0.579
SSR mean	0.276	0.442	0.606
INDEL mean	0.208	0.434	0.595

[a] *H. annuus*, *H. anomalus*, *H. argophyllus*, *H. deserticola*, *H. paradoxus*, and *H. tuberosus*

INDELs were equally polymorphic ($P = 0.12$–0.82), SSRs were equally polymorphic in exons and untranslated regions (UTRs) ($P = 0.20$–0.87), and amplicons spanning exons were as polymorphic as amplicons spanning exons and introns ($P = 0.16$–0.75) among the three germplasm groups (Table 2). Heterozygosity and SSR length were uncorrelated (correlations ranged from 0.10 to 0.14 for different germplasm groups).

Discussion

The EST-SSR and INDEL markers described here greatly increase the supply of highly portable DNA markers for comparative mapping and other genotyping applications in sunflower (Supplemental Tables 3, 4). Of the 1,829 SSR or INDEL markers developed for sunflower (Tang et al. 2002, 2003; Yu et al. 2002, 2003; Gandhi et al. 2005; Pashley et al. 2006), 30% target transcribed loci. Since our analyses were completed, the CGP has produced 257,833 additional sunflower ESTs, the number of sunflower ESTs deposited in public databases (DDBJ/EMBL/GenBank) has more than tripled, and the number of unigenes in the *H. annuus* transcript assembly has more than doubled (Kozik et al. 2002; http://cgpdb.ucdavis.edu/ cgpdb2/). SSRs ($n \geq 5$) were found in 10.9% of the 17,904 unigenes in the *H. annuus* transcript assembly screened in the present study. The latest transcript assemblies developed by the CGP harbor more than 4,000 SSRs ($n \geq 5$) and supply templates for EST-SSR marker development beyond those identified in the present study (Supplemental Table 1). Of the EST-SSR and INDEL markers screened in the present study, 55.1% were polymorphic in elite × elite and 83.3% were polymorphic in elite × exotic mapping populations (Tables 1, 2; Supplemental Table 4); hence, less than 10% of the transcribed loci in sunflower can be genetically mapped using SSRs. SNPs are significantly more abundant

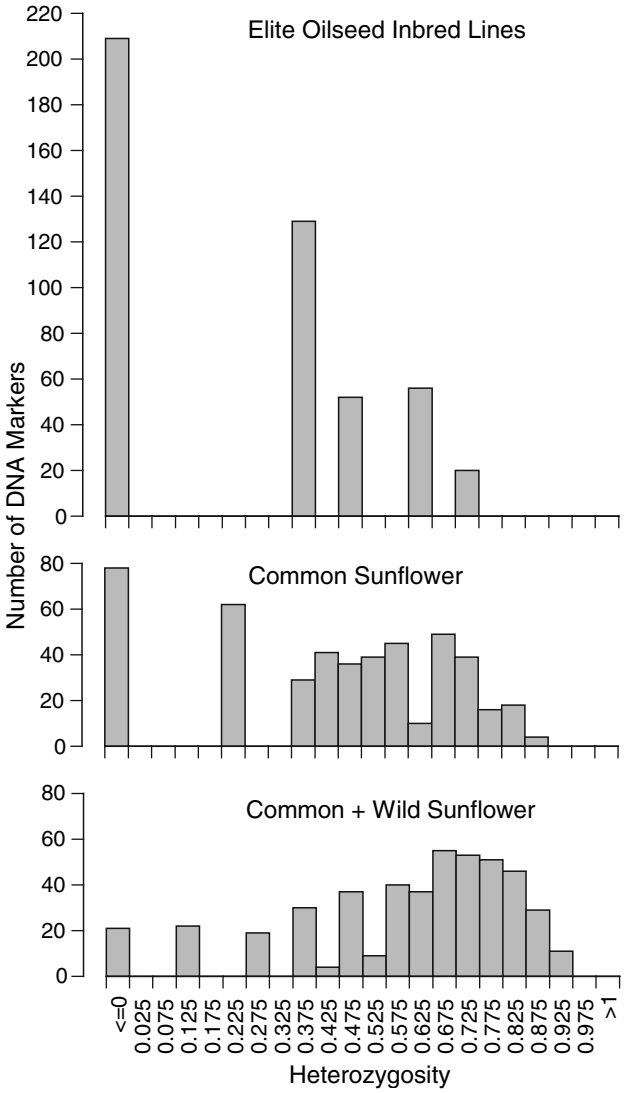

Fig. 2 Heterozygosity distributions for 466 common sunflower EST-SSR and INDEL markers among four elite oilseed inbred lines (elite), eight elite and exotic germplasm accessions of common sunflower (common), and wild germplasm accessions of *H. argophyllus*, *H. tuberosus*, *H. anomalus*, *H. paradoxus*, and *H. deserticola* (wild)

than SSRs and INDELs in eukaryotic genomes (Taramino and Tingey 1996; Bhattramakki et al. 2002; Rafalski 2002a, b) and have been discovered among elite alleles in a high frequency of the genes screened thus far in sunflower (Kolkman et al. 2004, 2007; Hass et al. 2006; Liu and Burke 2006; Schuppert et al. 2006; Tang et al. 2006). Kolkman et al. (2007) discovered SNPs in 80 of 81 loci by resequencing alleles from 10 elite inbred lines, whereas polymorphic SSRs were only found in 8 of the 81 loci. While SNPs are infrequent or lacking among elite alleles at some loci, they can nearly always be identified between elite and wild alleles (Kolkman et al. 2004, 2007; Liu and Burke 2006; Schuppert et al. 2006).

EST-SSR polymorphism trends in the present study paralleled those previously reported for non-genic SSRs among elite and exotic germplasm sources (Tang and Knapp 2003; Tang et al. 2003) and EST-SSRs in coding and non-coding sequences (Pashley et al. 2006). Pashley et al. (2006) found no significant difference in SSR polymorphisms between coding and non-coding sequences. Nor did we (Table 2). The four oilseed sunflower inbred lines screened in the present study (RHA801, RHA373, RHA377, and HA383) were previously screened for SSR length polymorphisms by Tang et al. (2003) using 300 non-genic SSR markers known a priori to be polymorphic among elite inbred lines. The mean heterozygosity for the latter among the four elite oilseed inbred lines was 0.379, which was 1.4-fold greater than the mean heterozygosity for the random sample of EST-SSR markers screened in the present study (0.267). Pashley et al. (2006) reported similar differences in heterozygosity between randomly selected samples of genic and non-genic SSRs (non-genic SSRs were 1.3-fold more polymorphic than EST-SSRs).

Most of the EST-SSR and INDEL markers developed for common sunflower amplify alleles from closely related sunflower species and should have broad utility for comparative mapping in *Helianthus* (Fig. 1; Pashley et al. 2006), and perhaps among closely related genera in tribe Heliantheae, e.g., *Echinacea*, *Parthenium*, and *Xanthium*. Cross-amplification percentages typically increase as phylogenetic distances decrease and are usually greater for primers complementary to coding than non-coding DNA sequences (Eujayl et al. 2004; Saha et al. 2004; Guo et al. 2006). Pashley et al. (2006) found common sunflower EST-SSR markers to be 1.5-fold more transferable to two wild sunflower species than non-genic SSR markers (73% of the EST-SSRs tested amplified alleles from *H. angustifolia* L. and *H. verticillatus* Small, as opposed to 50% for non-genic SSR markers). We found no difference, perhaps because the species we screened are close relatives of *H. annuus* (Rieseberg 1991; Rieseberg et al. 1991)—82% or more of the genic and non-genic SSR markers developed for common sunflower amplify alleles from one or more of the wild species screened (*H. petiolaris*, *H. paradoxus*, *H. anomalus*, *H. deserticola*, *H. argophyllus*, and *H. tuberosus*) (Fig. 1; Tang et al. 2002; Yu et al. 2002, 2003; Burke et al. 2004; Lai et al. 2005b).

DNA marker resources are limited for many species in the Compositae other than sunflower and lettuce (Kiers et al. 2000; Van Cutsem et al. 2003; Kim et al. 2004; Acquadro et al. 2005). We assessed the potential cross-taxa utility of sunflower EST-SSR and INDEL markers within the Compositae by screening species from two of the nine subfamilies and two of the 45 tribes—safflower from tribe Cardueae in the Carduoideae subfamily and lettuce and prickly lettuce from tribe Cichorieae in the Cichorioideae subfamily (Jansen et al. 1991; Funk et al. 2005; http://www.ncbi.nlm.nih.gov/Taxonomy/). Even though 69 sunflower EST-SSR markers amplified alleles from safflower, a species with limited DNA marker resources (Raina et al. 2005; Vilatersana et al. 2005), and 81 sunflower EST-SSR markers amplified alleles from *Lactuca*, a genus with significant DNA marker resources (Landry et al. 1987; Kesseli et al. 1994) (Fig. 1), cross-species utility was limited; 84–86% of the EST-SSR or INDEL markers failed to amplify safflower, lettuce, or prickly lettuce alleles. The development of DNA markers with broad applicability across genera in the Compositae has been challenging (Chapman et al. 2007), and RFLP analyses with heterologous probes, even from the closest eudicot clade (Solanaceae, Asterid I), have not enabled cross-family synteny analyses in sunflower (Compositae, Asterid II) (Chase et al. 1993; Fulton et al. 2002; Dominguez et al. 2003). Such analyses should be greatly facilitated by denser genetic mapping of transcribed loci, especially on a microsyntenic scale (Timms et al. 2006). The development of EST datatbases for 10 genera and 18 species, including safflower and chicory (*Cichorium intybus* L.), in four subfamilies (Asteroideae, Carduoideae, Cichorioideae, and Mutisioideae (http://cgpdb.ucdavis.edu), supplies a wealth of cDNA sequence templates for DNA marker development, comparative mapping, and other genotyping applications in the Compositae.

Acknowledgments This research was supported by grants from the National Science Foundation Plant Genome Program (No. 0421630) and United States Department of Agriculture Plant Genome Program (No. 2000-04292) to S.J.K, R.W.M, and L.H.R, and Advanta Seeds, Pioneer Hi-Bred International, Syngenta, and the Paul C. Berger Endowment to S.J.K.

References

Acquadro A, Portis E, Lee D, Donini P, Lanteri S (2005) Development and characterization of microsatellite markers in *Cynara cardunculus* L. Genome 48:217–225

Altschul SF, Gish W (1996) Local alignment statistics. Meth Enzymol 266:460–480

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Bhattramakki D, Dolan M, Hanafey M, Wineland R, Vaske D, Register JC, Tingey SV, Rafalski A (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

Ben C, Hewezi T, Jardinaud MF, Bena F, Ladouce N, Moretti S, Tamborindeguy C, Liboz T, Petitprez M, Gentzbittel L (2005) Comparative analysis of early embryonic sunflower cDNA libraries. Plant Mol Biol 57:255–270

Burke JM, Lai Z, Salmaso M, Nakazato T, Tang S, Heesacker A, Knapp SJ, Rieseberg LH (2004) Comparative mapping and rapid karyotypic evolution in the genus *Helianthus*. Genetics 167:449–457

Chapman MA, Chang J, Weisman D, Kesseli RV, Burke JM (2007) Universal markers for comparative mapping and phylogenetic analysis in the Asteraceae (Compositae). Theor Appl Genet 115:747–755

Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, Mishler BD, Duvall MR, Price RA, Hills HG, Qiu YL (1993) Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene rbcL. Ann Mo Bot Gard 80:528–580

Diatchenko L, Lau YF, Campbell AP, Chenchik A, Moqadam F, Huang B, Lukyanov S, Lukyanov K, Gurskaya N, Sverdlov ED, Siebert PD (1996) Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. Proc Natl Acad Sci USA 93:6025–6030

Dominguez I, Graziano E, Gebhardt C, Barakat A, Berry S, Arús P, Delseny M, Barnes S (2003) Plant genome archaeology: evidence for conserved ancestral chromosome segments in dicotyledonous plant species. Plant Biotechnol J 1:91–99

Eujayl I, Sledge MK, Wang L, May GD, Chekhovskiy K, Zwonitzer JC, Mian MAR (2004) *Medicago truncatula* EST-SSRs reveal cross-species genetic markers for *Medicago* spp. Theor Appl Genet 108:414–422

Fernandez P, Paniego N, Lew S, Hopp HE, Heinz RA (2003) Differential representation of sunflower ESTs in enriched organ-specific cDNA libraries in a small scale sequencing project. BMC Genomics 30:40–49

Frary A, Xu Y, Liu J, Mitchell S, Tedeschi E, Tanksley S (2005) Development of a set of PCR-based anchor markers encompassing the tomato genome and evaluation of their usefulness for genetics and breeding experiments. Theor Appl Genet 111:291–312

Fulton TM, Van der Hoeven R, Eannetta NT, Tanksley SD (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell 14:1457–1467

Funk VA, Bayer RJ, Keeley S, Chan R, Watson L, Gemeinholzer B, Schilling E, Panero JL, Baldwin BG, Garcia-Jagas N, Susanna A, Jansen RK (2005) Everywhere but Antarctica: using a supertree to understand the diversity and distribution of the Compositae. Biol Skr 55:343–374

Gandhi S, Heesacker AF, Freeman CA, Argyris J, Bradford K, Knapp SJ (2005) The self-incompatibility locus (*S*) and quantitative trait loci for self-pollination and seed dormancy in sunflower. Theor Appl Genet 111:619–629

Green P (1996) PHRAP (http://www.phrap.org)

Guo W, Wang W, Zhou B, Zhang T (2006) Cross-species transferability of *G. arboreum*-derived EST-SSRs in the diploid species of *Gossypium*. Theor Appl Genet 112:1573–1581

Han Z, Wang C, Song X, Guo W, Gou J, Li C, Chen X, Zhang T (2006) Characteristics, development and mapping of *Gossypium hirsutum* derived EST-SSRs in allotetraploid cotton. Theor Appl Genet 112:430–439

Hass CG, Tang S, Leonard S, Traber MG, Miller JF, Knapp SJ (2006) Three non-allelic epistatically interacting methyltransferase mutations produce novel tocopherol (vitamin E) profiles in sunflower. Theor Appl Genet 113:767–782

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877

Jansen RK, Michaels HJ, Palmer JD (1991) Phylogeny and character evolution in the Asteraceae based on chloroplast DNA restriction site mapping. Syst Bot 16:98–115

Kesseli RV, Paran I, Michelmore RW (1994) Analysis of a detailed genetic linkage map of *Lactuca sativa* (lettuce) constructed from RFLP and RAPD markers. Genetics 136:1435–1446

Kiers AM, Mes TH, van der Meijden R, Bachmann K (2000) A search for diagnostic AFLP markers in *Cichorium* species with emphasis on endive and chicory cultivar groups. Genome 43:470–476

Kim DH, Heber D, Still DW (2004) Genetic diversity of *Echinacea* species based upon amplified fragment length polymorphism markers. Genome 47:102–111

Kolkman J, Slabaugh MB, Bruniard J, Berry S, Bushman SB, Olungu C, Maes N, Abratti G, Zambelli A, Miller JF, Leon A, Knapp SJ (2004) Acetohydroxyacid synthase mutations conferring resistance to imidazolinone or sulfonylurea herbicides in sunflower. Theor Appl Genet 109:1147–1155

Kolkman JM, Berry ST, Leon AJ, Slabaugh MB, Tang S, Gao W, Shintani DK, Burke JM, Knapp SJ (2007) Single nucleotide polymorphisms and linkage disequilibrium in sunflower. Genetics 177:457–468

Kozik, A, Michelmore RW, Knapp SJ, Matvienko MS, Rieseberg L, Lin H, van Damme M, Lavelle D, Chevalier P, Ziegle J, Ellison P, Kolkman J, Slaubaugh MB, Livingston K, Zhou LZ, Church S, Edberg S, Jackson L, Bradford KJ (2002) Sunflower and lettuce ESTs developed by the Compositae Genome Program (http://cgpdb.ucdavis.edu/)

Lai Z, Livingstone K, Zou Y, Church SA, Knapp SJ, Andrews J, Rieseberg LH (2005a) Identification and mapping of SNPs from ESTs in sunflower. Theor Appl Genet 111:1532–1544

Lai Z, Nakazato T, Salmaso M, Burke JM, Tang S, Knapp SJ, Rieseberg LH (2005b) Extensive chromosomal repatterning and the evolution of sterility barriers in hybrid sunflower species. Genetics 171:291–303

Landry BS, Kesseli RV, Farrara B, Michelmore RW (1987) A genetic map of lettuce (*Lactuca sativa* L.) with restriction fragment length polymorphism, isozyme, disease resistance, and morphological markers. Genetics 116:331–337

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

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

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30:194–200

Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res 8:4321–4326

Ott J (1999) Analysis of human genetic linkage. Johns Hopkins Univ Press, Baltimore

Paniego N, Echaide M, Munoz M, Fernandez L, Torales S, Faccio P, Fuxan I, Carrera M, Zandomeni R, Suarez EY, Hopp HE (2002) Microsatellite isolation and characterization in sunflower (*Helianthus annuus* L.). Genome 45:34–43

Park YH, Alabady MS, Ulloa M, Sickler B, Wilkins TA, Yu J, Stelly DM, Kohel RJ, el-Shihy OM, Cantrell RG (2005) Genetic mapping of new cotton fiber loci using EST-derived microsatellites in an interspecific recombinant inbred line cotton population. Mol Genet Genomics 274:428–441

Pashley CH, Ellis JR, McCauley DE, Burke JM (2006) EST databases as a source for molecular markers: lessons from *Helianthus*. J Hered 97:381–388

Pawlowski K, Kunze R, de Vries S, Bisseling T (1994) Isolation of total, poly (A) and polysomal RNA from plant tissues. In: Gelvin SB, Schilperoort RA (eds) Plant molecular biology manual. Kluwer Academic Publishers, Norwell, pp 1–13

Peakall R, Gilmore S, Keys W, Morgante M, Rafalski A (1998) Cross-species amplification of soybean (*Glycine max*) simple sequence repeats (SSRs) within the genus and other legume genera: implications for the transferability of SSRs in plants. Mol Biol Evol 15:1275–1287

Poncet V, Rondeau M, Tranchant C, Cayrel A, Hamon S, de Kochko A, Hamon P (2006) SSR mining in coffee tree EST databases:

potential use of EST-SSRs as markers for the *Coffea* genus. Mol Genet Genomics 276:436–449

Rafalski A (2002a) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 5:94–100

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

Raina SN, Sharma S, Sasakuma T, Kishii M, Vaishnavi S (2005) Novel repeated DNA sequences in safflower (*Carthamus tintorius* L.) (Asteraceae): cloning, sequencing, and physical mapping by fluorescence in situ hybridization. J Hered 96: 424–429

Rieseberg LH (1991) Homoploid reticulate evolution in *Helianthus* (Asteraceae): evidence from ribosomal genes. Am J Bot 78:1218–1237

Rieseberg LH, Beckstrom-Sternberg SM, Liston A, Arias DM (1991) Phylogenetic and systematic inferences from chloroplast DNA and isozyme variation in *Helianthus* sect. *Helianthus* (Asteraceae). Syst Bot 16:50–76

Saha MC, Mian MA, Eujayl I, Zwonitzer JC, Wang L, May GD (2004) Tall fescue EST-SSR markers with transferability across several grass species. Theor Appl Genet 109:783–791

Schuppert GF, Tang S, Slabaugh MB, Knapp SJ (2006) The sunflower high-oleic mutant *Ol* carries variable tandem repeats of *FAD2–1*, a seed-specific oleoyl-phosphatidyl choline desaturase. Mol Breed 17:241–256

Tamborindeguy C, Ben C, Liboz T, Gentzbittel L (2004) Sequence evaluation of four specific cDNA libraries for developmental genomics of sunflower. Mol Genet Genomics 271:367–375

Tang S, Knapp SJ (2003) Microsatellites uncover extraordinary diversity in native American land races and wild populations of cultivated sunflower. Theor Appl Genet 106:990–1003

Tang S, Yu JK, Slabaugh MB, Shintani DK, Knapp SJ (2002) Simple sequence repeat map of the sunflower genome. Theor Appl Genet 105:1124–1136

Tang S, Kishore VK, Knapp SJ (2003) PCR-multiplexes for a genome-wide framework of simple sequence repeat marker loci in cultivated sunflower. Theor Appl Genet 107:6–19

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

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

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res 11:1441–1452

Thiel T, Michalek W, Varshney RK, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet 106:411–422

Timms L, Jimenez R, Chase M, Lavelle D, McHale L, Kozik A, Lai Z, Heesacker A, Knapp S, Rieseberg L, Michelmore R, Kesseli R (2006) Analyses of synteny between *Arabidopsis thaliana* and species in the Asteraceae reveal a complex network of small syntenic segments and major chromosomal rearrangements. Genetics 173:2227–2235

Van Cutsem P, du Jardin P, Boutte C, Beauwens T, Jacqmin S, Vekemans X (2003) Distinction between cultivated and wild chicory gene pools using AFLP markers. Theor Appl Genet 107:713–718

Varshney RK, Graner A, Sorrells ME (2005a) Genic microsatellite markers in plants: features and applications. Trends Biotech 23:48–55

Varshney RK, Sigmund R, Börner A, Korzun V, Stein N, Sorrells ME, Langridge P, Graner A (2005b) Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. Plant Sci 168:195–202

Vilatersana R, Garnatje T, Susanna A, Garcia-Jacas N (2005) Taxonomic problems in *Carthamus* (Asteraceae) RAPD markers and sectional classification. Bot J Linn Soc 147:375–383

Yu JK, Mangor J, Thompson L, Edwards KJ, Slabaugh MB, Knapp SJ (2002) Allelic diversity of simple sequence repeat markers among elite inbred lines in cultivated sunflower. Genome 45:652–660

Yu JK, Tang S, Slabaugh MB, Heesacker A, Cole G, Herring M, Soper J, Han F, Webb DM, Thompson L, Edwards KJ, Berry S, Leon A, Olungu C, Maes N, Knapp SJ (2003) Towards a saturated molecular genetic linkage map for sunflower. Crop Sci 43:367–387

Yu JK, Dake TM, Singh S, Benscher D, Li W, Gill B, Sorrells ME (2004a) Development and mapping of EST-derived simple sequence repeat markers for hexaploid wheat. Genome 47:805–818

Yu JK, La Rota M, Kantety RV, Sorrells ME (2004b) EST-derived SSR markers for comparative mapping in wheat and rice. Mol Genet Genomics 271:742–751