Journal of Heredity 2006:97(4):381–388
doi:10.1093/jhered/esl013
Advance Access publication July 13, 2006

EST Databases as a Source for Molecular Markers: Lessons from *Helianthus*

CATHERINE H. PASHLEY, JENNIFER R. ELLIS, DAVID E. MCCAULEY, AND JOHN M. BURKE

From the Department of Biological Sciences, Vanderbilt University, VU Station B 351634, Nashville, TN 37232.

Address correspondence to J. M. Burke at the Department of Plant Biology, University of Georgia, Athens, GA 30602, or e-mail: jmburke@uga.edu.

Abstract

Expressed sequence tag (EST) databases represent a potentially valuable resource for the development of molecular markers for use in evolutionary studies. Because EST-derived markers come from transcribed regions of the genome, they are likely to be conserved across a broader taxonomic range than are other sorts of markers. This paper describes a case study in which the publicly available cultivated sunflower (*Helianthus annuus*) EST database was used to develop simple sequence repeat (SSR) markers for use in the genetic analysis of a rare sunflower species, *Helianthus verticillatus*, as well as the more widespread *Helianthus angustifolius*. EST-derived SSRs were found to be more than 3 times as transferable across species as compared with anonymous SSRs (73% vs. 21%, respectively). Moreover, EST-SSRs whose primers were located within protein-coding sequence were more readily transferable than those derived from untranslated regions, and the former loci were no less variable than the latter. The utility of existing EST databases as a means for facilitating population genetic analyses in plants was further explored by cross-referencing publicly available EST resources against available lists of rare or invasive flowering plant taxa. This survey revealed that more than one-third of all plant-derived EST collections of sufficient size could conceivably serve as a source of EST-SSRs for the analysis of rare, endangered, or invasive plant species worldwide.

In recent years, simple sequence repeats (SSRs, a.k.a. microsatellites) have become the marker of choice for population genetic analyses. SSRs consist of tandem repeats of short (1–6 bp) nucleotide motifs (Gupta and others 1996), and these repetitive stretches are distributed throughout the genome, occurring both in protein-coding and noncoding regions (Toth and others 2000). Variation in SSR length occurs primarily due to slipped-strand mispairing during DNA replication (Toth and others 2000; Katti and others 2001; Li and others 2002), and mutations of this sort occur at a much higher frequency than do point mutations and insertions/deletions (Rossetto 2001). Thus, SSRs reveal much higher levels of polymorphism than do most other marker systems.

Although the utility of SSRs in population genetic studies is well established (Holton 2001), the isolation and characterization of such markers via traditional methods (i.e., the screening of size-fractionated genomic DNA libraries) are costly and time consuming (Squirrell and others 2003), making the de novo development of SSRs unrealistic for some taxa. In addition, the polymerase chain reaction (PCR) primers necessary for the amplification of such markers are frequently species specific. Thus, it is often difficult to make meaningful interspecific comparisons (e.g., comparing levels of genetic diversity between rare and common species or between species with different mating systems and/or dispersal strategies) using SSRs; indeed, the use of different sets of markers in different species confounds species differences with possible locus-specific effects, especially when relatively small numbers of loci are employed. Over the past decade, however, there has been a tremendous increase in the availability of DNA sequence data from a wide variety of taxa, including a wealth of expressed sequence tags (ESTs) that are typically unedited, automatically processed, single-pass sequences produced from cDNAs. Moreover, it has recently been shown that EST-based SSR markers (EST-SSRs) can be rapidly and inexpensively developed from existing EST databases (Gupta and others 2003; Bhat and others 2005). Thus, the use of such databases for marker development appears to be a promising alternative to the development of traditional "anonymous" SSRs following standard methods.

The primary limitation of using EST databases as a source of molecular markers is that this approach relies on existing genomic resources, and suitable databases are often only available to researchers who are studying economically important species. Adding to this difficulty is the fact that only a fraction of all ESTs contain an SSR. For example, it has been suggested that the frequency of SSR-containing sequences in plant-derived EST databases is typically on the order of 2–5% (Kantety and others 2002). Once developed, however, EST-SSRs are likely to be useful across a much broader taxonomic range than are anonymous SSRs owing

to the fact that the former come exclusively from transcribed regions of the genome. It is therefore possible that these markers will prove to be particularly valuable for the investigation of population genetic phenomena in close relatives of species with existing genomic resources.

One potential concern with regard to the use of EST-SSRs is that, because they are derived from genic regions, selection on these loci might influence the estimation of population genetic parameters. A recent study by Woodhead and others (2005) has, however, suggested that this may be a nonissue as estimates of population differentiation (i.e., F_{ST}) based on EST-SSRs are largely congruent with those based on anonymous SSRs. Moreover, the handful of large-scale studies that have been performed to date suggest that only a very small fraction of all genes have experienced recent positive selection (e.g., Tiffin and Hahn 2002; Clark and others 2003). Another more general concern with regard to using SSRs across taxonomic boundaries is that they might produce a higher level of null alleles. However, this is less likely to be a concern with EST-derived markers when compared with anonymous SSRs as the former reside in more conserved regions of the genome (Liewlaksaneeyanawin and others 2004). Consistent with this view, null alleles have been found to be less of a problem for EST-SSRs as compared with anonymous SSRs in cross-taxon applications in a variety of cases (e.g., Leigh and others 2003; Rungis and others 2004).

Although reports on the transferability of EST-SSRs have become increasingly common, particularly in plants (e.g., Decroocq and others 2003; Thiel and others 2003; Varshney and others 2005), the majority of these studies have demonstrated transferability between economically important taxa (e.g., domesticated crops) or from such taxa to their close relatives that have been targeted for their potential use in breeding programs (Cordeiro and others 2001; Saha and others 2004). To date, very few studies have directly compared the transferability of EST versus anonymous SSRs in a common set of taxa (Chagne and others 2004; Liewlaksaneeyanawin and others 2004; Gutierrez and others 2005). The present study focuses on a somewhat underappreciated yet clearly important application of EST-SSRs—the development of markers from existing EST databases for use in evolutionary and/or conservation genetic studies in related taxa. This paper provides a case study of the utility of freely available EST resources for the development of markers necessary for the genetic analysis of a rare sunflower species, *Helianthus verticillatus* Small, and the closely related, but much more common, species *Helianthus angustifolius* L. More specifically, a novel suite of 48 polymorphic SSRs developed from the cultivated sunflower (*Helianthus annuus* L.) EST database are described, and their transferability and levels of polymorphisms are compared with those of a suite of anonymous SSRs that were also developed from *H. annuus*. In addition, the likely utility of this approach as a means for facilitating conservation genetic analyses in general is explored by summarizing publicly available EST resources and cross-referencing these results against available lists of rare or invasive flowering plant taxa.

Materials and Methods

Study System

The genus *Helianthus* is a member of Compositae (a.k.a. the Asteraceae), which is one of the largest and most diverse flowering plant families, comprising approximately one-tenth of all known angiosperm species. *Helianthus* is composed of 51 species, with 14 annuals and 37 perennials. The annual species are divided into 3 sections, the majority of which (12 of 14) come from the section *Helianthus* (Seiler and Gulya 2004). Included in section *Helianthus* is the cultivated sunflower (*H. annuus*) that is one of the world's most important oilseed crops and also a major source of confectionery seeds and ornamental plants (Putt 1997). Over the past several years, cultivated sunflower has been the subject of a major EST sequencing effort. Indeed, the Compositae Genome Project Database (CGPDB) (http://cgpdb.ucdavis.edu) now contains approximately 44 000 cultivated sunflower EST sequences corresponding to approximately 19 000 unique genes. Of these, 2360 have been found to contain at least one SSR.

The perennial species of *Helianthus* are divided into section *Ciliares*, which is subdivided into 2 series, and section *Atrorubens*, which is subdivided into 4 series (Seiler and Gulya 2004). Of particular interest for the present paper are *H. verticillatus* (sect. *Atrorubens*, ser. *Corona-solis*; EE Schilling personal communication) and *H. angustifolius* (sect. *Atrorubens*, ser. *Angustifolii*; Seiler and Gulya 2004). The whorled sunflower *H. verticillatus* is a diploid ($n = 17$), perennial, and rhizomatous plant that is named for the fact that its leaves are borne in whorls of 3 or 4 below the inflorescence. The range of this species, which is a candidate for federal listing as an endangered species, is limited to the southeastern United States of America, where it known from only 3 locations (one population each in Alabama, Georgia, and Tennessee). Although there has been some disagreement regarding the origin of *H. verticillatus*, with some authors suggesting that it might simply represent aberrant individuals derived from hybridization (Beatley 1963; Heiser and others 1969), perhaps between *Helianthus grosseserratus* Martens and *H. angustifolius* L. (Heiser and others 1969), recent data suggest that *H. verticillatus* is most likely a "genuine" species of nonhybrid origin (Matthews and others 2002; Ellis and others 2006). As with *H. verticillatus*, *H. angustifolius* is a diploid perennial (Seiler and Gulya 2004); unlike *H. verticillatus*, however, *H. angustifolius* is a common plant that can be found throughout much of the eastern United States of America.

Plant Materials and DNA Extraction

Seeds of common sunflower were obtained from the North Central Regional Plant Introduction Station (NCRPIS, Ames, IA), nicked with a razor blade, allowed to germinate on moistened filter paper, and then transplanted and grown in the Vanderbilt University Department of Biological Sciences greenhouses. Sampled accessions included Arizona (NCRPIS accession: Ames 14400), Arkansas (PI 613727), California (PI 613732), Colorado (PI 586840), Iowa (PI 597895), Nebraska

(PI 586865), North Dakota (PI 586810), Ohio (Ames 23238), Texas (Ames 7442), Utah (PI 531009), Washington (PI 531016), and Wyoming (PI 586822). One sample per accession was then selected for DNA extraction. Total genomic DNA was isolated from 200 mg of fresh leaf tissue using the DNeasy plant mini kit (Qiagen, Valencia, CA). For *H. verticillatus*, leaf samples were collected from individuals in the 3 known wild populations (Cherokee County, AL; Floyd County, GA; and Madison County, TN), whereas leaf samples of *H. angustifolius* were collected from 2 sites (Cherokee County, AL, and Madison County, TN). DNA was extracted from 4 individuals from each population of *H. verticillatus* and 6 individuals from each population of *H. angustifolius* using a modification of the Doyle JJ and Doyle JL (1987) cetyl trimethyl ammonium bromide method. All DNA samples were quantified using a TKO-100 fluorometer (Hoefer Scientific Instruments, San Francisco, CA).

EST-SSR Development and Analysis

Sunflower EST sequences from the CGPDB (http://cgpdb.ucdavis.edu) were scanned with a PERL script designed to identify SSR-containing sequences consisting of di-, tri-, and tetranucleotide repeats with a minimum of 5, 4, or 3 subunits, respectively. Primer pairs flanking repeats were designed using PRIMER3 (freely available at http://www.broad.mit.edu/genome_software/), and these primers were then tested against the 12 wild *H. annuus* individuals listed above. Out of 188 primer pairs that were found to successfully amplify *H. annuus* DNA, an arbitrary set of 48 markers that amplified consistently across individuals and produced scorable polymorphisms were selected for inclusion in this study. Specific information for each locus, including CGPDB contig ID, repeat motif, and primer sequences, is available on request.

SSR genotyping was performed using a modified version of the fluorescent labeling protocol of Schuelke (2000), as detailed in Wills and others (2005). Reactions were performed in 20 μl total volume containing 10 ng of template DNA for all taxa except *H. verticillatus*, where 2 ng was used; 30 mM Tricine, pH 8.4 KOH; 50 mM KCl; 2 mM $MgCl_2$; 125 μM of each deoxynucleoside triphosphate; 0.2 μM M13 Forward (−29) sequencing primer labeled with VIC, 6FAM, or TET; 0.2 μM reverse primer; 0.04 μM forward primer; and 2 units of *Taq* polymerase. PCR was performed on a PTC-100 Peltier Thermal Cycler (MJ Research, South San Francisco, CA) using a touchdown (Don and others 1991) regimen as follows: 3 min at 95 °C; 10 cycles of 30 s at 94 °C, 30 s at 65 °C, and 45 s at 72 °C, annealing temperature decreasing to 55 °C by 1 °C per cycle; followed by 30 cycles of 30 s at 94 °C, 30 s at 55 °C, 45 s at 72 °C, followed by 20 min at 72 °C. Amplification products were visualized on an MJ Research BaseStation automated DNA sequencer. MapMarker® 1000 ROX size standards (BioVentures Inc., Murfreesboro, TN) were run in each lane to allow for accurate determination of fragment size, and alleles were called using the software package CARTOGRAPHER (MJ Research).

Anonymous SSR Analysis

Forty-eight anonymous SSRs developed by Tang and others (2003) were arbitrarily selected for analysis. These markers are known to be polymorphic across *H. annuus* and have already been mapped in cultivated sunflower (Tang and others 2003). Amplifications were performed as described for the EST-SSRs, except that the forward primer from each pair was directly labeled. Thus, the forward primer concentration was increased to 0.2 μM, and the labeled M13 Forward (−29) sequencing primer was eliminated.

Assessing Transferability for EST-Derived and Anonymous SSRs

All 96 primer pairs (corresponding to the 48 EST-derived loci and the 48 anonymous loci) were initially screened for transferability across taxa by attempting amplification on a panel consisting of 2 individuals each of *H. annuus*, *H. angustifolius* and *H. verticillatus*, as well as 1 individual each of *Lactuca sativa* L. cv. Salinas and *Lactuca serriola* L. (DNA kindly provided by R. W. Michelmore, University of California, Davis). The latter 2 species, which occur in a different subfamily within the Compositae, were included to assess the level of marker transferability across greater evolutionary distances within the family. Amplification products were visualized on 2% agarose gels stained with ethidium bromide, and only those primer pairs for which both *H. verticillatus* and *H. angustifolius* yielded strong amplification products of similar size to that of *H. annuus* were selected for further analysis.

The primer pairs selected from the initial cross-taxon screen were used to genotype the 12 *H. angustifolius* and *H. verticillatus* individuals. The resulting fluorescent profiles for each locus were then assigned a quality score (Q) from 1 to 5 as per Leigh and others (2003), where 1 denotes a PCR product of expected size, with a clear signal and no stuttering; 2 denotes a scorable product accompanied by faint stutter bands; 3 denotes an unscorable ladder of stutter bands or a multilocus amplification product; 4 denotes a weak/unreliable product; and 5 denotes a failed amplification. For each locus with $Q = 1$ or 2, genetic diversity in each species was estimated as $H_e = 1 - \sum p_i^2$ (where p_i is the frequency of the ith allele at the locus). These calculations ignored the possibility of null alleles, and missing data were excluded.

Results

Development of EST-SSRs

As noted above, 48 polymorphic *H. annuus* EST-SSRs were arbitrarily selected for inclusion in the later stages of this study. This collection of loci included di-, tri-, and tetranucleotide repeats. Trinucleotide motifs were the most abundant (22 of 48 loci), followed by tetranucleotide (18 of 48 loci) and dinucleotide repeats (1 of 48 loci). The remaining 7 primer pairs flanked multiple repeat motifs, with 3 flanking a pair of trinucleotide repeat motifs, 2 flanking a pair of tetranucleotide motifs, 1 flanking 3 trinucleotide motifs, and

Table 1. Summary of amplification results of EST-SSRs and anonymous SSRs from *Helianthus annuus* to *Helianthus verticillatus* and *Helianthus angustifolius* when visualized via agarose gel electrophoresis

Result	EST-SSRs	Anonymous SSRs
Strong products of expected size in both taxa	35	24
Weak products of expected size in both taxa	3	1
Strong product in *H. verticillatus* but not in *H. angustifolius*	4	6
Strong product in *H. angustifolius* but not in *H. verticillatus*	2	2
Product outside of *H. annuus* size range	1	4
No amplification in most/all individuals of either taxa	3	11

1 flanking both a tri- and a tetranucleotide motif. The single most common motif amplified was ATG (5 of 48 loci). The number of alleles per locus varied from 2 to 11 ($A = 4.96 \pm 0.268$, mean ± standard error [SE]) in the panel of 12 wild *H. annuus* individuals, and levels of genetic diversity ranged from 0.469 to 0.855 ($H_e = 0.66 \pm 0.017$). Overall, 39 of 48 primer pairs amplified products of the expected length. The exceptions produced products ranging from approximately 50 to 500 bases larger than expected and were most likely due to the presence of introns, which are not present in EST sequences.

Cross-Species Transferability of SSRs within the Genus *Helianthus*

Thirty-five of the 48 (73%) EST-SSR primer pairs tested produced a strong amplification product of the expected size in both *H. verticillatus* and *H. angustifolius*, whereas only 24 of 48 (50%) of the anonymous markers did so (Table 1). The difference is significant ($\chi^2 = 5.32$, df = 1, $P < 0.05$). Three EST-SSRs and one anonymous SSR produced weak amplification products of the expected size in both taxa. One EST-SSR and 4 anonymous loci produced amplification products outside of the expected size range, whereas 2 EST-SSRs and 6 anonymous SSRs failed to consistently produce an amplification product in either taxon. The remaining primer pairs (8 EST derived and 17 anonymous) appeared to work better in either *H. verticillatus* or *H. angustifolius*. Overall, 39 EST-SSRs and 30 anonymous SSRs produced strong amplification products of the expected size in *H. verticillatus*, whereas 36 EST-SSRs and 26 anonymous SSRs did so in *H. angustifolius*.

Marker Transferability across the Compositae

None of the EST-derived or anonymous primer pairs amplified a strong product of the expected size in both species of lettuce analyzed. Two of the EST-derived primer pairs produced strong amplicons of the expected size from *L. sativa* DNA, whereas 2 of the anonymous SSRs did so from *L. serriola* DNA. All other primer pairs failed to produce an amplification product, produced weak amplification products, produced amplicons that were >100 bp outside of the expected range, or produced multilocus amplification profiles.

Table 2. Comparison of mean heterozygosities between markers that were/were not transferable among species within the genus *Helianthus*. All estimates are derived from *H. annuus*

Marker type	Status	Mean	SE	*P*
EST-SSRs	Transferable	0.643	0.020	0.21
	Nontransferable	0.691	0.032	
Anonymous SSRs	Transferable	0.858	0.011	0.42
	Nontransferable	0.872	0.012	

Comparison of Marker Types for Genotyping across the Genus *Helianthus*

As noted above, 35 of the 48 EST-SSRs and 24 of the 48 anonymous SSRs showed promise for genotyping in both *H. verticillatus* and *H. angustifolius* (Table 1). These markers were next used to genotype a panel of 12 individuals from each of *H. verticillatus* and *H. angustifolius*. A quality score (*Q*) of 1 or 2 when visualized via acrylamide gel electrophoresis is indicative of a reliable marker, and all 35 of the apparently transferable EST-SSRs (73% of the original 48) fell into these 2 categories for both *H. verticillatus* and *H. angustifolius* (data not shown). In contrast, only 10 anonymous SSRs (21% of the original 48) fell into these 2 categories for both *H. verticillatus* and *H. angustifolius*. The difference was highly significant ($\chi^2 = 31.67$, df = 1, $P < 0.001$). For both marker types, the primer pairs that could be successfully transferred across taxa revealed slightly, but not significantly ($P > 0.20$), less variation in *H. annuus* than was present at the nontransferable loci (Table 2).

The number of alleles per locus in *H. verticillatus* ranged from 1 to 9 for both the EST-SSRs (3.34 ± 0.389, mean ± SE) and the anonymous SSRs (3.85 ± 0.608), and heterozygosities ranged from 0 to 0.855 (0.427 ± 0.048) and from 0 to 0.868 (0.502 ± 0.077), respectively. For *H. angustifolius*, the number of alleles per locus varied from 1 to 7 (2.23 ± 0.256) for the EST-SSRs and from 2 to 7 (3.86 ± 0.430) for the anonymous SSRs, whereas heterozygosities ranged from 0 to 0.781 (0.234 ± 0.044) and from 0.180 to 0.826 (0.572 ± 0.051), respectively. Polymorphism data for *H. annuus* is available for all 35 of the transferable EST-SSR primer pairs, whereas comparable data are available for only 13 of the 24 transferable anonymous primers (Tang and Knapp 2003). The number of alleles per locus in *H. annuus* ranged from 2 to 11 (4.94 ± 0.330) for the EST-SSRs and from 6 to 15 (10.69 ± 0.523) for the anonymous SSRs, and heterozygosities ranged from 0.469 to 0.855 (0.64 ± 0.020) and from 0.780 to 0.915 (0.858 ± 0.009), respectively. Of the 35 transferable EST-SSR primer pairs that were polymorphic in *H. annuus*, 6 were found to be monomorphic within *H. verticillatus*, whereas 16 were monomorphic

Table 3. Comparison of levels of genetic variation detected by the 2 different marker types for each of the 3 *Helianthus* species

Taxon	Marker type	No. polymorphic loci	H_e^a			A_P		
			Mean	SE	P^b	Mean	SE	P^b
H. annuus	EST	48	0.66	0.014	<0.0001	4.96	0.302	<0.0001
	Anonymous	25	0.86	0.020		11.12	0.418	
H. verticillatus	EST	28	0.53	0.041	ns	3.93	0.412	ns
	Anonymous	12	0.54	0.063		4.08	0.630	
H. angustifolius	EST	19	0.43	0.044	ns	3.26	0.338	ns
	Anonymous	14	0.57	0.051		3.86	0.394	

ns, not significant.

[a] Estimates based on data from polymorphic loci.

[b] Adjusted for multiple comparisons (Rice 1989).

within *H. angustifolius*. In contrast, only 1 of the 13 anonymous SSRs that was transferable to *H. verticillatus* was monomorphic, and none of the 14 anonymous SSRs that could be scored in *H. angustifolius* were monomorphic in that taxon. Excluding the monomorphic loci, the average heterozygosity for *H. verticillatus* and *H. angustifolius* was 0.53 ± 0.040 and 0.43 ± 0.045, respectively, when amplified with EST-SSR markers and 0.54 ± 0.073 and 0.57 ± 0.053, respectively, when amplified with anonymous markers. Although the anonymous SSRs revealed higher levels of polymorphism than did the EST-SSRs for all 3 *Helianthus* taxa (Table 3), the differences were significant for only *H. annuus*. It is important to note here, however, that the *H. annuus* anonymous SSR polymorphism data come from a different study (Tang and Knapp 2003). Thus, although the geographic range covered and sample sizes employed were similar, these data come from a different set of individuals than do the EST-SSR data. The overall levels of diversity across taxa were generally positively correlated for both marker types, both in terms of the number of alleles per locus and in terms of heterozygosity, although this was only statistically significant for the EST-SSR allele number between *H. verticillatus* and *H. angustifolius* ($P < 0.05$) and for both EST-SSR allele number ($P < 0.001$) and heterozygosity ($P < 0.001$) between *H. annuus* and *H. verticillatus*.

SSR Motif and Origin of EST Microsatellite Sequences

For all EST-SSRs, ESTSCAN2 (Iseli and others 1999; Lottaz and others 2003; http://www.ch.embnet.org/software/ESTScan2.html) was used to determine whether the priming sites and repeat motifs were located in protein-coding sequence or in untranslated regions (UTRs). For 16 of the 48 loci under consideration, both of the priming sites (and thus the associated repeat motifs) appear to be located within coding sequence, whereas both primers (along with the associated repeat motifs) appear to be located in an UTR in 8 cases. The remaining 24 loci had one primer each in coding and UTR sequence, and of these loci with "split" priming sites, the SSRs themselves were located in coding sequence in 7 cases and in an UTR in 17. Not surprisingly, loci for which both primers were found in coding sequence were significantly more transferable than those for which one or both primers were located in the UTRs (100% vs. 60% transferability, respectively; Fisher's exact test, $P = 0.002$). With regard to levels of polymorphism, however, it is noteworthy that SSRs that were located within coding sequence were no less polymorphic (measured as heterozygosity) than were those that were located in an UTR (0.68 ± 0.029 vs. 0.64 ± 0.021, respectively; Wilcoxon rank sum test, $P = 0.36$).

As noted in Materials and Methods, the 48 EST-SSR primer pairs used in this study originated from a suite of 188 *H. annuus* EST-SSR markers. Of these, just more than 50% (95/188) had microsatellites that were located in coding sequence. When used to genotype the 12 *H. annuus* accessions used in this study, heterozygosities in this larger set of markers ranged from 0 to 0.851 (0.463 ± 0.031) for SSRs in coding sequence and 0 to 0.885 (0.453 ± 0.030) for those from UTRs. Approximately a fifth of these markers (39/188) were monomorphic, and those located in protein-coding sequence were no more likely to be monomorphic than those located in UTRs ($\chi^2 = 0.216$, df = 1, $P = 0.642$).

Discussion

Marker Transferability across the Genus *Helianthus*

Our results indicate that EST-SSRs are significantly more transferable across *Helianthus* species than are anonymous SSRs, with 73% of the *H. annuus* EST-SSRs surveyed being transferable to both *H. verticillatus* and *H. angustifolius* as compared with just 21% of the anonymous SSRs. Note that these species represent 2 of the most divergent sections within the genus *Helianthus* (Schilling 1997). Not surprisingly, the most readily transferable EST-SSRs were those whose primers were located in coding sequence, presumably due to greater sequence conservation in such regions. A more surprising result was that SSRs located in coding regions were no less variable than those located in UTRs, perhaps owing to a bias toward trinucleotide repeats in these regions. In fact, a general trend toward the occurrence of trinucleotide repeats in open reading frames has been noted in EST databases derived from a number of plant taxa, including wheat (Gupta and others 2003), barrel medic (Eujayl and others 2004), tall

fescue (Saha and others 2004), pine (Chagne and others 2004), and barley (Toth and others 2000), and the higher level of variability of these repeat motifs likely results from the fact that the gain/loss of trinucleotide repeat units has no effect on the protein-coding frame. As such, novel trinucleotide length variants are less likely to be selected against as compared with di- or tetranucleotide variants.

Although the high transferability of EST-SSRs reported here is in general agreement with the results of previous studies in other taxa, very few of these studies have actually made direct comparisons using the 2 different marker types on a common set of individuals (but see Chagne and others 2004; Liewlaksaneeyanawin and others 2004; Gutierrez and others 2005). More often, these papers have concentrated on simply documenting the fact that EST-SSRs can often be transferred across taxa (e.g., Decroocq and others 2003; Thiel and others 2003; Varshney and others 2005) or on comparing the levels of diversity revealed by EST-SSRs and anonymous SSRs within a single species. As has been reported in a variety of other taxa, for example, sugarcane (Cordeiro and others 2001), rice (Cho and others 2000), wheat (Eujayl and others 2002; Leigh and others 2003), and pine (Chagne and others 2004; Liewlaksaneeyanawin and others 2004), EST-SSRs in *Helianthus* are generally less polymorphic than anonymous SSRs. Although the increased frequency of monomorphism in the nonsource taxa (especially *H. angustifolius*) suggests that there may be trade-offs when using EST-SSRs, these markers were far from invariant overall, with estimates of gene diversity averaging 0.53 across taxa (Table 3). This value is considerably higher than the values commonly associated with studies using allozymes (Hamrick and Godt 1990, 1996), suggesting that EST-SSRs are likely to reveal ample levels of polymorphism for most population genetic applications.

Beyond exhibiting a much higher level of cross-species transferability, the EST-SSRs surveyed in this study also produced consistently "cleaner" results than did the anonymous SSRs. Indeed, of the 35 EST-SSRs that could be successfully transferred to *H. verticillatus* and *H. angustifolius*, the majority (60% and 80%, respectively) produced a clear amplification product with no evidence of stutter ($Q = 1$). In contrast, only one anonymous SSR proved to be of equivalent quality in *H. verticillatus* and *H. angustifolius*; although the balance were scorable, they were all accompanied by stutter banding ($Q = 2$). The general superiority of EST-SSRs in terms of data quality has been previously noted by several other authors (Eujayl and others 2001, 2002; Leigh and others 2003; Woodhead and others 2005) and may be a by-product of shorter repeat motifs in genic regions as opposed to elsewhere in the genome and/or the aforementioned tendency toward trinucleotide repeats (Liepelt and others 2001).

Our results suggest that the cultivated sunflower EST database will be a rich resource for the development of SSRs for use across the genus *Helianthus*. In this context, it is worth noting that *Helianthus* contains a number of other species of potential conservation interest, including both threatened/endangered species (e.g., *Helianthus paradoxus* Heiser, *Helianthus eggertii* Small, and *Helianthus schweintzii* T. and G.) and naturalized weeds (e.g., *Helianthus tuberosus* L., *Helianthus petiolaris* Nutt., and *Helianthus ciliaris* DC.; Seiler and Gulya 2004). In view of our findings, the best strategy for exploiting the sunflower EST database for use in these taxa would be to preferentially target repeat motifs that are located within protein-coding sequence. Assuming that the primers flanking such repeats are themselves found within the protein-coding region, these loci should exhibit the highest level of cross-species transferability while being no less variable than SSRs found in UTRs. In fact, this tendency for EST-SSRs to have similar levels of average diversity regardless of their locations (i.e., open reading frames vs. UTRs) has also been documented in pines (Chagne and others 2004), suggesting that the targeting of repeat motifs located in protein-coding sequence may be a generally useful strategy.

EST Databases as a Source of Molecular Markers

As outlined in the introduction, one of the rate-limiting steps in population genetic analyses is often the development of markers for use in a new study system. Moreover, many marker systems are more or less species specific, which can complicate comparisons of the level and organization of population genetic variation across taxa. When the generally high transferability of EST-SSRs is combined with the fact that most population genetic analyses rely on a rather small number of markers (e.g., Richards and others 2004; Vornam and others 2004; Szczys and others 2005), it seems likely that even modest EST collections could provide a way around these difficulties for researchers interested in studying the relatives of taxa from which these resources are derived. In fact, an estimated 2–5% of all plant-derived ESTs are thought to harbor SSRs (Kantety and others 2002), meaning that EST databases containing as few as 500 sequences could conceivably provide enough candidate SSRs to facilitate population genetic studies. What remains less clear to evolutionary and/or conservation biologists is the extent to which existing EST resources overlap with species of interest.

As of March 2005, the National Center for Biotechnology Information EST database (http://www.ncbi.nlm.nih.gov/dbEST/) housed EST collections consisting of $\geq$500 sequences for 160 flowering plants comprising 100 unique genera from 30 different orders. In order to gauge the utility of these sequences as a source of markers for use in conservation-related studies, we cross-referenced these EST collections against the US Fish and Wildlife Service's threatened and endangered species database (http://www.fws.gov/endangered/wildlife.html), the 2004 International Union for Conservation of Nature and Natural Resources Red List of threatened species (http://www.iucnredlist.org/), and the US State and Federal Composite List of Noxious Weeds (http://plants.usda.gov/).

After accounting for overlap across lists, it appears that more than one-third of all plant-derived EST collections containing a minimum of 500 sequences (corresponding to 37 of the 100 genera with such resources) could potentially serve as a source of EST-SSRs for the analysis of rare, endangered, or invasive plants species worldwide (data available on

request). Note that this is most likely a somewhat conservative estimate as 1) our survey was primarily based on data from US agencies, although the most critically endangered species worldwide were also included, and 2) only those EST collections that were derived from a congener of the focal species were included in the tally, yet EST-SSRs are often transferable across greater taxonomic distances. Moreover, whereas rare and invasive plants were chosen to illustrate the likely utility of existing EST resources as a source for molecular markers, these resources have the potential to facilitate population genetic research in a much wider variety of other taxa. Of course, the utility of any particular EST collection for marker development will need to be assessed on a case-by-case basis.

Acknowledgments

This work was supported by grants from the National Science Foundation Plant Genome Research Program (DBI-0332411 to J.M.B.) and the National Research Initiative of the US Department of Agriculture (USDA) Cooperative State Research, Education and Extension Service (#03-35300-13104 to J.M.B.). Mark Chapman, Aizhong Liu, Jessica Wenzler, David Wills, and 3 anonymous reviewers provided comments on an earlier version of the manuscript. *Helianthus annuus* sequence data were obtained from The Compositae Genome Project Web site (http://compgenomics.ucdavis.edu), which was supported by the USDA Initiative for Future Agriculture and Food Systems program.

References

Beatley JC. 1963. The sunflowers (genus *Helianthus*) in Tennessee. J Tenn Acad Sci 38:135–54.

Bhat PR, Krishnakumar V, Hendre PS, Rajendrakumar P, Varshney RK, Aggarwal RK. 2005. Identification and characterization of expressed sequence tags-derived simple sequence repeats, markers from robusta coffee variety 'C×R' (an interspecific hybrid of *Coffea canephora* × *Coffea congensis*). Mol Ecol Notes 5:80–3.

Chagne D, Chaumeil P, Ramboer A, Collada C, Guevara A, Cervera MT, Vendramin GG, Garcia V, Frigerio JMM, Echt C, Richardson T, Plomion C. 2004. Cross-species transferability and mapping of genomic and cDNA SSRs in pines. Theor Appl Genet 109:1204–14.

Cho YG, Ishii T, Temnykh S, Chen X, Lipovich L, McCouch SR, Park WD, Ayres N, Cartinhour S. 2000. Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa* L.). Theor Appl Genet 100:713–22.

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, Ferriera S, Wang G, Zheng X, White TJ, Sninsky JJ, Adams MD, Cargill M. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302: 1960–3.

Cordeiro GM, Casu R, McIntyre CL, Manners JM, Henry RJ. 2001. Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to erianthus and sorghum. Plant Sci 160:1115–23.

Decroocq V, Fave MG, Hagen L, Bordenave L, Decroocq S. 2003. Development and transferability of apricot and grape EST microsatellite markers across taxa. Theor Appl Genet 106:912–22.

Don RH, Cox PT, Wainwright BJ, Baker K, Mattick JS. 1991. Touchdown PCR to circumvent spurious priming during gene amplification. Nucleic Acids Res 19:4008.

Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull 19:11–5.

Ellis JR, Pashley CH, Burke JM, McCauley DE. 2006. High genetic diversity in a rare and endangered sunflower as compared to a common congener. Mol Ecol. Forthcoming.

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 108:414–22.

Eujayl I, Sorrells M, Baum M, Wolters P, Powell W. 2001. Assessment of genotypic variation among cultivated durum wheat based on EST-SSRS and genomic SSRS. Euphytica 119:39–43.

Eujayl I, Sorrells ME, Baum M, Wolters P, Powell W. 2002. Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. Theor Appl Genet 104:399–407.

Gupta PK, Balyan IS, Sharma PC, Ramesh B. 1996. Microsatellites in plants: a new class of molecular markers. Curr Sci 70:45–54.

Gupta PK, Rustgi S, Sharma S, Singh R, Kumar N, Balyan HS. 2003. Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. Mol Genet Genomics 270:315–23.

Gutierrez MV, Patto MCV, Huguet T, Cubero JI, Moreno MT, Torres AM. 2005. Cross-species amplification of *Medicago truncatula* microsatellites across three major pulse crops. Theor Appl Genet 110:1210–7.

Hamrick JL, Godt MJW. 1990. Allozyme diversity in plant species. In: Brown AHD, Clegg MT, Kahler AL, Weir BS, editors. Plant population genetics, breeding and genetic resources. Sunderland, MA: Sinauer Associates, Inc. p 43–63.

Hamrick JL, Godt MJW. 1996. Effects of life history traits on genetic diversity in plant species. Philos Trans R Soc Lond Ser B Biol Sci 351: 1291–8.

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

Holton TA. 2001. Plant genotyping by analysis of microsatellites. In: Henry RJ, editor. Plant genotyping: the DNA fingerprinting of plants. Wallingford, CT: CABI Publishing. p 15–28.

Iseli C, Jongeneel CV, Bucher P. 1999. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol:138–48.

Kantety RV, Rotal ML, Matthews DE, Sorrells ME. 2002. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol Biol 48:501–10.

Katti MV, Ranjekar PK, Gupta VS. 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol 18:1161–7.

Leigh F, Lea V, Law J, Wolters P, Powell W, Donini P. 2003. Assessment of EST- and genomic microsatellite markers for variety discrimination and genetic diversity studies in wheat. Euphytica 133:359–66.

Li YC, Korol AB, Fahima T, Beiles A, Nevo E. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol Ecol 11:2453–65.

Liepelt S, Kuhlenkamp V, Anzidei M, Vendramin GG, Ziegenhagen B. 2001. Pitfalls in determining size homoplasy of microsatellite loci. Mol Ecol Notes 1:332–5.

Liewlaksaneeyanawin C, Ritland CE, El-Kassaby YA, Ritland K. 2004. Single-copy, species-transferable microsatellite markers developed from loblolly pine ESTs. Theor Appl Genet 109:361–9.

Lottaz C, Iseli C, Jongeneel CV, Bucher P. 2003. Modeling sequencing errors by combining Hidden Markov models. Bioinformatics 19:ii103–12.

Matthews JF, Allison JR, Ware SRT, Nordman C. 2002. *Helianthus verticillatus* Small (Asteraceae) rediscovered and redescribed. Castanea 67:13–24.

Putt ED. 1997. Early history of sunflower. In: Scheiter AA, editor. Sunflower production and technology. Madison, WI: American Society of Agronomy. p 1–19.

Rice WR. 1989. Analyzing tables of statistical tests. Evolution 43: 223–5.

Richards CM, Reilley A, Touchell D, Antolin MF, Walters C. 2004. Microsatellite primers for Texas wild rice (*Zizania texana*), and a preliminary test of the impact of cryogenic storage on allele frequency at these loci. Conserv Genet 5:853–9.

Rossetto M. 2001. Sourcing of SSR markers from related plant species. In: Henry RJ, editor. Plant genotyping: the DNA fingerprinting of plants. Wallingford, CT: CABI Publishing. p 211–24.

Rungis D, Berube Y, Zhang J, Ralph S, Ritland CE, Ellis BE, Douglas C, Bohlmann J, Ritland K. 2004. Robust simple sequence repeat markers for spruce (*Picea* spp.) from expressed sequence tags. Theor Appl Genet 109:1283–94.

Saha MC, Mian MAR, Eujayl I, Zwonitzer JC, Wang LJ, May GD. 2004. Tall fescue EST-SSR markers with transferability across several grass species. Theor Appl Genet 109:783–91.

Schilling EE. 1997. Phylogenetic analysis of *Helianthus* (Asteraceae) based on chloroplast DNA restriction site data. Theor Appl Genet 94:925–33.

Schuelke M. 2000. An economic method for the fluorescent labeling of PCR fragments. Nat Biotechnol 18:233–4.

Seiler GJ, Gulya TJ. 2004. Exploration for wild *Helianthus* species in North America: challenges and opportunities in the search for global treasures. In: Seiler GJ, editor. Proceedings of the 16th International Sunflower Conference; Fargo, ND. Paris: International Sunflower Association, p 43–68.

Squirrell J, Hollingsworth PM, Woodhead M, Russell J, Lowe AJ, Gibby M, Powell W. 2003. How much effort is required to isolate nuclear microsatellites from plants? Mol Ecol 12:1339–48.

Szczys P, Hughes CR, Kesseli RV. 2005. Novel microsatellite markers used to determine the population genetic structure of the endangered Roseate Tern, *Sterna dougallii*, in Northwest Atlantic and Western Australia. Conserv Genet 6:461–6.

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

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 106:411–22.

Tiffin P, Hahn MW. 2002. Coding sequence divergence between two closely related plant species: *Arabidopsis thaliana* and *Brassica rapa* ssp. *pekinensis*. J Mol Evol 54:746–53.

Toth G, Gaspari Z, Jurka J. 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 10:967–81.

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

Vornam B, Decarli N, Gailing O. 2004. Spatial distribution of genetic variation in a natural beech stand (*Fagus sylvatica* L.) based on microsatellite markers. Conserv Genet 5:561–70.

Wills DM, Hester ML, Liu AZ, Burke JM. 2005. Chloroplast SSR polymorphisms in the Compositae and the mode of organeller inheritance in *Helianthus annuus*. Theor Appl Genet 110:941–7.

Woodhead M, Russell J, Squirrell J, Hollingsworth PM, Mackenzie K, Gibby M, Powell W. 2005. Comparative analysis of population genetic structure in *Athyrium distentifolium* (Pteridophyta) using AFLPs and SSRs from anonymous and transcribed gene regions. Mol Ecol 14:1681–95.

Received October 25, 2005
Accepted June 5, 2006

Corresponding Editor: James Hamrick