American Journal of Botany

American Journal of Botany 99(2): 209–218. 2012.

Genomics of Compositae weeds: EST libraries, microarrays, and evidence of introgression[1]

Zhao Lai[2], Nolan C. Kane[3], Alex Kozik[4], Kathryn A. Hodgins[3], Katrina M. Dlugosch[5], Michael S. Barker[5], Marta Matvienko[4], Qian Yu[2], Kathryn G. Turner[3], Stephanie Anne Pearl[6], Graeme D. M. Bell[3], Yi Zou[2], Chris Grassa[3], Alessia Guggisberg[3], Keith L. Adams[3], James V. Anderson[7], David P. Horvath[7], Richard V. Kesseli[8], John M. Burke[6], Richard W. Michelmore[4], and Loren H. Rieseberg[2,3,9]

[2]Department of Biology and Center for Genomics and Bioinformatics, Indiana University, Bloomington, Indiana 47405 USA; [3]Department of Botany and Biodiversity Research Centre, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada; [4]Department of Plant Sciences and Genome Center, University of California, Davis, California 95616 USA; [5]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721 USA; [6]Department of Plant Biology, University of Georgia, Athens, Georgia 30602 USA; [7]Biosciences Research Laboratory, USDA-ARS, 1605 Albrecht Boulevard, Fargo, North Dakota 58105-5674 USA; and [8]Biology Department, University of Massachusetts, Boston, Massachusetts, USA

- *Premise of study:* Weeds cause considerable environmental and economic damage. However, genomic characterization of weeds has lagged behind that of model plants and crop species. Here we describe the development of genomic tools and resources for 11 weeds from the Compositae family that will serve as a basis for subsequent population and comparative genomic analyses. Because hybridization has been suggested as a stimulus for the evolution of invasiveness, we also analyze these genomic data for evidence of hybridization.
- *Methods:* We generated 22 expressed sequence tag (EST) libraries for the 11 targeted weeds using Sanger, 454, and Illumina sequencing, compared the coverage and quality of sequence assemblies, and developed NimbleGen microarrays for expression analyses in five taxa. When possible, we also compared the distributions of Ks values between orthologs of congeneric taxa to detect and quantify hybridization and introgression.
- *Results:* Gene discovery was enhanced by sequencing from multiple tissues, normalization of cDNA libraries, and especially greater sequencing depth. However, assemblies from short sequence reads sometimes failed to resolve close paralogs. Substantial introgression was detected in *Centaurea* and *Helianthus,* but not in *Ambrosia* and *Lactuca*.
- *Conclusions:* Transcriptome sequencing using next-generation platforms has greatly reduced the cost of genomic studies of nonmodel organisms, and the ESTs and microarrays reported here will accelerate evolutionary and molecular investigations of Compositae weeds. Our study also shows how ortholog comparisons can be used to approximately estimate the genome-wide extent of introgression and to identify genes that have been exchanged between hybridizing taxa.

Key words: Asteraceae; Compositae; ESTs; hybridization; introgression; invasive plants; microarray development; next-generation sequencing; sequence assemblies; transcriptome sequencing; weeds.

Weedy and invasive plants cause considerable damage to the economy and environment. In North America, direct economic costs due to production losses in agriculture and forestry, as well as the cost of control measures, are estimated at $30–40 billion annually (Pimentel et al., 2000, 2005; Myers and Bazely, 2003; Colautti et al., 2006). Environmental costs are more difficult to estimate monetarily, but they can be profound and include extinction of species, loss of biodiversity, and degradation of ecosystem services (Wilcove et al., 1998; Simberloff, 2005; Dextrase and Mandrak, 2006). While the ecological and financial damage caused by weeds has stimulated considerable research into the ecology and evolution of weeds and invasive plants, genomic characterization has lagged behind that of model plants or crop species (Stewart et al., 2009).

Recent discussions of weed genomics have argued for the development of one or several model "weed" species that would be suitable for answering questions about weed biology and would serve to concentrate funding and intellectual efforts (Basu et al., 2004; Chao et al., 2005; Stewart et al., 2009). However, the community of scientists who study weedy and invasive plants is diverse, as are the traits that apparently contribute to the success of weedy species. This diversity is to be expected given that weeds are often broadly defined as "plants that grow

[1]Manuscript received 7 July 2011; revision accepted 16 September 2011.

The authors thank F. Bretagnolle, P. van Dijk, T. Gulya, J. Hierra, R. Hufbauer, L. Kiss, P. Kotanen, Y. Sapir, D. Lavelle, and G. Seiler for assistance in obtaining seed or tissue collections; and M. Stewart, M. Scascitelli, H. Luton, and K. Nurkowski for technical assistance. They also thank the sequencing and bioinformatics teams at the Joint Genome Institute, Genome Quebec, Indiana University's Center for Genomics and Bioinformatics and the David H. Murdock Research Institute for assistance with the generation and processing of the raw sequence data. National Science Foundation Awards 0421630 and 0820451 and Natural Sciences and Engineering Research Council of Canada Award 353026 provided funding.

[9]Author for correspondence (e-mail: lriesebe@mail.ubc.ca)

doi:10.3732/ajb.1100313

in disturbed areas" (Heiser, 2003). Thus, it may not be appropriate to focus the efforts of the community of weed and invasive plant researchers on one or even a handful of weeds. The advent of "next-generation," high-throughput sequencing technologies enables the identification of genetic changes that are frequently associated with the evolution of weedy and invasive plants, as well as those that are idiosyncratic. Such comparative genomic approaches exploit the diversity of weeds and invasive plants to answer questions about the ecological, evolutionary, and molecular mechanisms contributing to their successes.

The Compositae (Asteraceae) family is especially well suited for comparative genomic studies of weed evolution. The Compositae is one of the largest and most successful families of flowering plants (Stevens, 2001), with close to 24 000 named species that thrive in a great diversity of habitats, including some of the world's most inhospitable. Although the Compositae family contains several hundred economically valuable species (Dempewolf et al., 2008), it is perhaps best known for its noxious weeds such as thistles, knapweeds, ragweeds, and dandelions. Indeed, the Compositae includes eight of the 20 worst weeds in North America (Rice, 2011). Also, 36 of 181 North American species that have been newly introduced and are potentially invasive in Europe come from the Compositae (Forman, 2003). While the traits associated with successful Compositae weeds vary across taxa (Muth and Pigliucci, 2006), herbicide resistance (Peng et al., 2010) and growth-defense trade-offs (Mayrose et al., 2011) are commonly observed in weedy species in the family.

The two most economically important genera of the Compositae (*Helianthus* and *Lactuca*) are particularly interesting and complementary with regard to their reciprocal histories of domestication and the evolution of invasiveness. Sunflower was domesticated in North America, yet today 11 of the 49 species in the genus *Helianthus* (including *H. annuus*) are considered naturalized or invasive in Europe (DAISIE, 2009). Also, due to high levels of gene flow between cultivated and weedy sunflowers (Arias and Rieseberg, 1994; Linder et al., 1998), sunflower has been featured in debates about the role of crop–wild gene flow and transgene escape in the evolution of "super weeds" (Burke and Rieseberg, 2003; Ellstrand, 2003; Snow et al., 2003; Baack et al., 2008; Dechaine et al., 2009, 2010). Conversely, lettuce was domesticated in the Mediterranean region, yet today *Lactuca serriola* L. (the progenitor of cultivated lettuce) and 11 other wild species of *Lactuca* have become established in North America (Lebeda et al., 2004).

Here we report on the development of genomic tools and resources for 11 Compositae weeds (Table 1): *Ambrosia artemisiifolia* L. (common ragweed), *Ambrosia trifida* L. (giant ragweed), *Centaurea diffusa* Lam. (diffuse knapweed), *Centaurea stoebe* subsp. *micranthos* L. (spotted knapweed), *Centaurea solstitialis* L. (yellow starthistle), *Cirsium arvense* (L.) Scop. (Canada thistle), *Carthamus oxyacanthus* M. Bieb. (jeweled distaff thistle), *Helianthus annuus* L. (common sunflower), *Helianthus ciliaris* DC. (Texas blueweed), *L. serriola* (prickly lettuce), and *Taraxacum officinale* F. H. Wigg (dandelion). Most of these taxa are native to Europe or Central Asia and are invasive in North America or elsewhere. However, three of the targeted weeds have a reciprocal history of invasion: common and giant ragweed and common sunflower are native to North America and naturalized elsewhere (Heiser et al., 1969; Genton et al., 2005). While the majority of the target weeds are diploid, outcrossing annuals, there are several exceptions. Prickly lettuce is a selfing annual or biennial (Alexander, 2010). Canada thistle is an outcrossing perennial (Lalonde and Roitberg, 1994). Spotted knapweed and blueweed are perennials and have multiple ploidy levels (Heiser et al., 1969; Blair and Hufbauer, 2010). Dandelions are perennial, have multiple ploidy levels, and produce asexual seeds through apomixis (Verduijn et al., 2004).

While all of these weeds are able to colonize disturbed habitats such as cropland, abandoned fields, roadsides, and railroads, they vary in competitive ability and in their damage to the environment and to human health. Dandelion is a major lawn weed across the temperate world (Pimentel et al., 2000; Pimentel et al., 2005). Knapweeds and thistles are rangeland weeds that have colonized and degraded millions of hectares of pastures and rangeland in western North America (LeJeune and Seastedt, 2001). Ragweeds are abundant colonizers of disturbed habitats across much of temperate North America, Europe, Asia, and Australia, and allergens produced by their pollen are the primary cause of hay fever (Laaidi et al., 2003; Chauvel et al., 2006). Hay fever costs $3.5 billion per year in the United States alone in direct medical expenditures (Storms et al., 1997) and more than 10 times as much in lost workplace productivity (Lamb et al., 2006).

The genomic tools and resources that we describe here are intended to serve as the basis for subsequent population and comparative genomic analyses. In addition, we report on the coverage of 454, Illumina, and Sanger cDNA libraries and compare the quality of the assemblies from data generated with these three platforms. Last, we describe evidence that hybridization is associated with the evolution of several of the weeds investigated here and provide a preliminary report on the kinds of genes that appear to have been exchanged between the hybridizing taxa.

MATERIALS AND METHODS

***EST library development*—**Express sequence tag (EST) libraries were developed for one or more accessions of the 11 Compositae weeds targeted by this study (Table 1). RNA was isolated from a variety of tissues (Table 2) using either Trizol reagent (Invitrogen, Carlsbad, California, USA) or RNeasy Maxi (or Mini) kits (Qiagen, Valencia, California, USA), or a combination of the two methods. In the combined approach, the standard Trizol protocol was followed through the chloroform extraction step, then 0.53× volumes of 100% ethanol was added to the aqueous phase, the entire RNA/ethanol mixture was then applied to an RNeasy Maxi (or Mini) column, and the Qiagen protocol followed thereafter. Approximately equal amounts of total RNA isolated from each tissue type were pooled prior to EST library preparation.

Several different methods were used to generate EST libraries as sequencing technologies advanced (Table 2). For Sanger sequencing, we prepared standard libraries using the SMART (Clontech, Palo Alto, California, USA) approach or normalized libraries with the TRIMMER-DIRECT cDNA Normalization Kit (Evrogen, Moscow, Russia). The cDNA samples from both the standard and normalized EST libraries were size-fractionated through agarose gels into three classes (0.5–1 kb, 1–2 kb, and 2–3 kb) to reduce biases due to size during the subsequent cloning and sequencing steps.

For 454 sequencing (454 Life Sciences, Branford, Connecticut, USA), we employed modified oligo-dT primers during cDNA synthesis to reduce the length of mononucleotide runs associated with the poly(A) tail of mRNA. Mononucleotide runs reduce sequence quality and quantity due to excessive light production and crosstalk between neighboring cells. For common ragweed, we used a "broken chain" short oligo-dT primer to prime the poly(A) tail of mRNA during first strand cDNA synthesis (Meyer et al., 2009). cDNA was amplified and normalized with the TRIMMER-DIRECT cDNA Normalization Kit as above. Then normalized cDNA was prepared for sequencing following the standard genomic DNA shotgun protocol recommended by 454 Life

TABLE 1. Provenance information for Compositae weeds targeted in this study.

Taxon	Common name	Collection locality	Collection ID
Ambrosia artemisiifolia L.	Common ragweed	Biatorbagy, Hungary (latitude 47.46, longitude 18.81)	HU1-11
		Russell, MN, USA (latitude 44.19, longitude −95.57)	AA8-20
Ambrosia trifida L.	Giant ragweed	Jilin, Jilin, China (latitude 43.50, longitude 126.32	GNV8ASA01
		Dengta, Liaoning, China (latitude 41.25, longitude 123. 20)	GNV8ASA02
		Kampsville, IL, USA (latitude 39.16, longitude 90.37)	GNV8ASA03
		Bloomington, IN, USA (latitude 39.92, longitude 86.31)	GNV8ASA04
Carthamus oxyacanthus M. Bieb.	Jeweled distaff thistle	38 km north of crossroad just west of Mardin, Turkey	PI 407602
Centaurea diffusa Lam.	Diffuse knapweed	Kirklarelì, Turkey (latitude 41.45, longitude 27.14)	DK TR001-1L
		Roosevelt, WA, USA (latitude 45.44, longitude −120.12)	DK US022-31E
Centaurea stoebe subsp. *micranthos* L.	Spotted knapweed	Tetraploid genotype, Boston, MA, USA (latitude 42.29, longitude −71.04).	R. Kesseli - Cema #1A
Centaurea solstitialis L.	Yellow starthistle	Walnut Creek, CA, USA (latitude 37.95, longitude −122.05)	R. Kesseli - Ceso JH1 #1
		Santa Rosa, Argentina (latitude −37.39, longitude −64.08).	AR-13-24
Cirsium arvense (L.) Scop.	Canada thistle	Female plant, Fargo, ND, USA (latitude 46.93, longitude −96.86)	NW-22-1-M
		Male plant, Richmond Hill, ON, Canada (latitude 43.95, longitude −79.56)	Hodgins KN-ON
		Female plant, Lugoj, Romania (latitude 45.65, longitude 21.95)	Guggisberg, Bretagnolle & Zeltner 280808-2
Helianthus annuus L.	Common sunflower	Ma'ayan Tzvi, Israel (latitude 32.33, longitude 34.56)	Hann - ISI
		Port Augusta, Australia (latitude 32.29, longitude 137.47)	SAW3, USDA PI 653594
Helianthus ciliaris DC.	Texas blueweed	Tetraploid genotype, weed garden of the New Mexico State University Plant Science Research Center in Dona Ana County, NM, USA	L. Rieseberg - Hcil 1411
Lactuca serriola L.	Prickly lettuce	Davis, CA, USA	US96UC23
Taraxacum officinale F. H. Wigg	Dandelion	Apomictic triploid, Heteren, The Netherlands	P.J. van Dijk - Taof A68

Sciences. For cDNA synthesis of the other libraries, we either used the broken chain short oligo-dT primer described above or two different modified oligo-dT primers: one to prime the poly(A) tail of mRNA during first strand cDNA synthesis and another to further break down the stretches of poly(A) sequence during second strand cDNA synthesis (Beldade et al., 2006). We then normalized and amplified the cDNA using the TRIMMER-DIRECT cDNA Normalization Kit as above. After normalization, cDNA was fragmented to 500- to 800-bp fragments by either sonication or nebulization and size-selected to remove

TABLE 2. Plant tissues employed for EST library development.

Taxon	Collection ID	Roots	Leaves	Flower buds	Mature flowers	Fruits or seeds	Seedlings	Library / Sequence type [a]
Ambrosia artemisiifolia	HU1-11		x					N, SS / 454
	AA8-20		x					N, SS / 454
Ambrosia trifida	GNV8ASA01	x	x	x	x			N, DS / 454
	GNV8ASA02	x	x	x	x			N, DS / 454
	GNV8ASA03	x	x	x	x			N, DS / 454
	GNV8ASA04	x	x	x	x			N, DS / 454
Carthamus oxyacanthus	PI 407602	x	x		x			N, DS / 454
Centaurea diffusa	DK TR001-1L		x					N, DS / 454
	DK US022-31E		x					N, DS / 454
Centaurea stoebe subsp. *micranthos*	Cema #1A	x	x	x	x	x		N, SF / Sanger
Centaurea solstitialis	Ceso JH1 #1	x	x	x	x	x		N, SF / Sanger
	AR-13-24		x					N, DS / 454
Cirsium arvense [b]	NW-22-1-M	x	x	x	x			N, DS / 454
	Hodgins KN-ON		x					S / Illumina
	Guggisberg et al., 280808-2		x					S / Illumina
Helianthus annuus [c]	Several cultivars	x	x	x	x	x		N / Sanger
	Hann ISI						x	N, DS / 454
	SAW3						x	N / Illumina
Helianthus ciliaris	Hcil 1411	x	x	x	x			N, SF / Sanger
Lactuca serriola	US96UC23	x	x	x	x	x		N / Sanger
	US96UC23	x	x	x	x	x		N / Illumina
Taraxacum officinale [d]	Taof A68	x	x	x	x	x		S, SF/ Sanger

[a] N = normalized library; S = standard library; DS = double-stranded libraries; SS, SF = size-fractionated library.

[b] Illumina EST libraries for *Cirsium arvense* were generated as part of an analysis of gene regulatory evolution and are described in G. Bell et al. (unpublished manuscript)

[c] Sanger EST libraries for *H. annuus* previously described by Heesacker et al. (2008).

[d] Leaves of *Taraxacum officinale* were from plants sprayed with salicylic acid (4 mmol/L in 0.1% Triton X-100) or jasmonic acid (50 mmol/L in 0.1% Triton X-100) to induce defense-related gene expression.

small fragments using AMpure SPRI beads (Angencourt, Beverly, Massachusetts, USA). Then the fragmented ends were polished and ligated with adaptors. The optimal ligation products were selectively amplified and subjected to two rounds of size selection including gel electrophoresis and AMpure SPRI bead purification (Lai et al., 2011).

For Illumina sequencing, we prepared standard libraries using the mRNA-Seq (Illumina, San Diego, California, USA) approach or normalized libraries using customized approaches. For *L. serriola*, cDNA was synthesized using the mRNA-Seq cDNA Synthesis Kit (Illumina) prior to normalization with the TRIMMER-DIRECT cDNA Normalization Kit (M. Matvienko et al., unpublished). For the remaining libraries sequenced with Illumina (Table 2), cDNA was synthesized using the SMART PCR cDNA Synthesis Kit (Clontech, Palo Alto, California, USA) and then normalized with the TRIMMER-DIRECT Kit. The normalized libraries were then prepared for sequencing as recommended by Illumina. After determination of fragment size distributions on a Bioanalyzer (Agilent Technologies, Santa Clara, California, USA) and of concentrations with PicoGreen (Invitrogen), libraries were diluted for real-time quantitative PCR and sequenced.

***Processing and assembly of EST libraries*—**The Sanger EST libraries were sequenced using ABI 3730 machines (Life Technologies, Carlsbad, California, USA) at the Joint Genome Institute in Walnut Creek, California. Phred base calling, masking, trimming, and CAP3 assemblies (Huang and Madan, 1999) were conducted using the CGPdb bioinformatic pipelines (http://compgenomics.ucdavis.edu/index.php?link=tools; Compositae Genome Project, Genome Center, University of California-Davis). While the present paper provides the first published description of the development of these libraries and the accessions employed, the Sanger ESTs reported here were previously included in assemblies reported by Barker et al. (2008).

The 454 EST libraries were sequenced on GS-FLX machines (454 Life Sciences) at the Indiana University Center for Genomics and Bioinformatics (http://cgb.indiana.edu/), the David H. Murdock Research Institute (DHMRI; http://www.dhmri.org/about.html), or Genome Quebec (http://www.genomequebec.com/v2009/home/) using the standard 454 Titanium chemistry. The 454 sequences were cleaned using the program SnoWhite version 1.1.4 (http://evopipes.net/snowhite.html) (Barker et al., 2010) or the program ESTclean (http://sourceforge.net/projects/estclean/). Cleaned sequences were initially assembled with the program MIRA version 3.0 (Chevreux et al., 2004), using the "accurate,est,denovo,454" assembly mode. However, because in our experience, MIRA can be too aggressive in splitting up contigs with high coverage, we took the MIRA contigs and singletons and reassembled them with the program CAP3 at 94% identity (Huang and Madan, 1999).

The Illumina EST libraries were sequenced on Illumina GAII machines at the UC Davis Genome Center (http://www.genomecenter.ucdavis.edu/), Indiana University Center for Genomics and Bioinformatics, or at DHMRI. Illumina data were cleaned with customized scripts (http://code.google.com/p/atgc-illumina/) and assembled with the program CLC (http://www.clcdenovo.com/, CLC bio, Aarhus, Denmark) using the default settings or the program Trinity (http://trinityrnaseq.sourceforge.net/) using the Butterfly parameters –bfly_opts "–edge-thr=0.05 -V 5" to increase its ability to distinguish close paralogs.

Coverage offered by each of the assemblies was evaluated in terms of the number of unigenes, assembly length, and the proportion of ultra-conserved orthologs (UCOs) detected (Tables 3, 4) using the NCBI program blastx and an e-value threshold of 1e-10. The UCOs are 357 single-copy genes that are shared by *Arabidopsis thaliana,* humans, mice, yeast, fruit flies, and *Caenorhabditis elegans* (Kozik et al., 2008). Assembly quality was evaluated by analyzing the proportion of recently duplicated paralogs in the assembly, as well as the percentage of UCOs with full-length transcripts. The proportion of recently duplicated paralogs was determined by analyzing duplicate gene age distributions using the DupPipe (Barker et al., 2010) pipeline described in Barker et al. (2008). The rationale for this analysis is that assemblies of short reads or overaggressive assemblies may fail to distinguish between recently diverged paralogs. The percentage of full-length transcripts was determined using the UCO hits, where transcripts were considered full-length if they covered greater than 80% of the annotated UCO protein and included start and stop codons.

***Detection of hybridization*—**For several genera (*Ambrosia*, *Centaurea*, *Helianthus*, *Lactuca*), we have EST libraries from multiple taxa that frequently co-occur and potentially hybridize. To test for hybridization, we identified orthologs between all congeneric taxa using reciprocal best hits, as in Kane et al. (2009). The distribution of Ks values (number of synonymous substitutions per synonymous site) for orthologs should be centered around a Ks value corresponding to the time since the most recent common ancestor of the taxa involved. However, a secondary peak at a lower Ks value can be attributed to more recent gene flow (Wang and Hey, 2010). We identified significant peaks in the ortholog Ks distribution using SiZer (Chaudhuri and Marron, 1999). The number of significant peaks in the range $0 < Ks < 0.1$ was inferred with the maximum-likelihood approach in the EMMIX (McLachlan et al., 1999) package. The optimal number of peaks was inferred as the model that minimizes the Bayesian information criterion (BIC).

Gene Ontology (GO) categorization was performed on the genes found in introgressed and nonintrogressed peaks from the EMMIX analysis, using blastx searches with an e-value threshold of 10^{-10} against TAIR10 proteins (http://www.arabidopsis.org/). We tested for differences in GO annotations using χ^2 tests with P values computed from 100 000 Monte Carlo simulations in the program R (R Development Core Team, 2008). Major contributors to significant χ^2 tests ($P < 0.05$) were identified as in Barker et al. (2008), using residuals with absolute values greater than 2.

***Microarray development*—**In addition to the analysis of hybridization, we employed the EST libraries from six taxa (common ragweed, diffuse knapweed, spotted knapweed, yellow starthistle, Canada thistle, and common sunflower) to develop high-density expression microarrays in collaboration with Roche NimbleGen (Madison, Wisconsin, USA) (Table 5). The microarrays were developed to investigate expression differences associated with the evolution of weedy and invasive genotypes in different Compositae weeds. However, they should be useful for a wide range of ecological, evolutionary, and molecular studies of Compositae weeds and their wild relatives.

The NimbleGen high-density customized expression microarray service offers transcript-based probe design with long, isothermal probes. After the masking of repetitive elements, 2 or 3 unique probes were designed per unigene, with the remaining space on the array (usually less then 5%) filled with random probes for background correction. Both 4-plex and 12-plex expression microarray platforms were developed (Table 5).The platforms differ in the number of hybridizations that can be performed per array (4 vs. 12), as well as the number of probes per plex (72 000 vs. 135 000).

For common ragweed, diffuse knapweed, and Canada thistle, 12-plex arrays were developed using the transcriptome of a genotype from the invasive range of each species. The numbers of probes and unigenes chosen for array development are given in Table 5. Unigenes were mainly chosen for inclusion based on the quality, length, and uniqueness of sequence, but for ragweed we enriched slightly for stress-related transcripts.

For yellow starthistle, we developed a 4-plex array based on 24 545 unigenes from an invasive genotype of yellow starthistle and 9798 unigenes from an invasive genotype of spotted knapweed (Table 5). Two probes were chosen per contig.

Last, for common sunflower, we developed both 4-plex and 12-plex arrays (Table 5). The 4-plex expression array was based on a Sanger transcriptome assembly of cultivated sunflower ESTs, whereas the 12-plex expression array was based on the 454 titanium transcriptome assembly from a weedy genotype collected outside of the native range of the species.

***Databases*—**The National Center for Biotechnology Information (NCBI) recently announced that it might discontinue its Sequence Read Archive (SRA) and Trace Archive repositories for high-throughput sequencing data and that only assemblies will be archived in the future. This news is troubling because access to the raw reads will be needed for many studies in population and evolutionary genomics. Therefore, if the SRA is not continued at the NCBI or elsewhere, both the raw data and assemblies generated by this study will be archived on the Compositae Genome Project Database (http://compgenomics.ucdavis.edu/). The raw Sanger reads as well as the reference assemblies for all 22 EST libraries have already been submitted to and are accessible from GenBank and/or Dryad (see Table 6 for details).

RESULTS AND DISCUSSION

***EST sequencing*—**The sequencing of EST libraries provides a relatively inexpensive means for sampling transcribed genes from any given tissue or organism. As a consequence, EST sequencing has been the primary entry point for genomic studies of nonmodel organisms (Bouck and Vision, 2007; Vera et al.,

TABLE 3. ESTs and assembly statistics for Compositae weeds targeted by this study.

Taxon	Collection ID	Sequence type [a]	No. reads	Total sequence (Mbp)	No. unigenes	Total assembly length (Mbp)	% UCOs [b]	% Full-length transcripts [c]	% Paralogs with Ks < 0.1
Ambrosia artemisiifolia	HU1-11	454	701 460	185	71 179	38	85	8.8	40
	AA8-20	454	616 318	157	62 936	33	83	7.8	36
Ambrosia trifida	GNV8ASA01	454	609 298	221	57 285	38	91	21.6	50
	GNV8ASA02	454	238 943	95	28 574	18	77	22.5	44
	GNV8ASA03	454	206 343	81	25 378	16	72	23.5	40
	GNV8ASA04	454	192 795	77	25 120	16	76	25.2	35
Carthamus oxyacanthus	PI 407602	454	406 005	125	27 255	40	85	28.6	38
Centaurea diffusa	DK TR001-iL	454	407 817	183	48 936	31	77	20.2	67
	DK US022-31E	454	631 874	308	61 749	43	86	23.5	63
Centaurea stoebe subsp. *micranthos*	Cema #1A	Sanger	39 957	29	20 922	17	80	27.2	24
Centaurea solstitialis	Ceso JH1 #1	Sanger	40 406	30.5	22 917	19	79	26.8	26
	AR-13-24	454	649 880	274	43 503	32	92	30.6	56
Cirsium arvense	NW-22-1-M	454	3 770 510	1411	66 269	61	99	23.7	66
	Hodgins KN-ON	Illumina	39 316 660	2988	54 718	30	97	16.9	1.1
	Guggisberg et al., 280808-2	Illumina	39 411 764	2995	46 807	25	98	20.2	1.1
Helianthus annuus	Hann ISI	454	1 132 254	446	54 124	37	93	54.8	56
	SAW-3	Illumina	10 630 366	1063	37 108	13	83	13.7	0.9
	Several cultivars	Sanger	93 428	47.9	31 605	18	66	20.4	32
Helianthus ciliaris	Hcil 1411	Sanger	21 589	16.6	14 857	12	68	26.8	22
Lactuca serriola	US96UC23	Sanger	55 452	34.2	19 877	14	67	24.4	32
	US96UC23	Illumina	91 048 987	7539	66 733	61	100	43.8	1
Taraxacum officinale	Taof A68	Sanger	41 278	29	15 761	12	56	34.4	41

[a] Sanger assemblies previously reported in Barker et al. (2008)

[b] Percentage of ultra-conserved orthologs (UCOs) found in EST library. UCOs refer to 357 single-copy genes that are shared by *Arabidopsis thaliana,* humans, mice, yeast, fruit flies, and *Caenorhabditis elegans* (Kozik et al., 2008).

[c] Percentage of full-length transcripts calculated for UCOs.

2008). EST sequence data have a broad array of applications ranging from gene discovery and annotation (Sterky et al., 2004; Albert et al., 2005), to molecular marker development (Lai et al., 2005; Ellis and Burke, 2007; Heesacker et al., 2008), to gene expression analyses, whether directly through sequencing (Simon et al., 2009) or indirectly through microarray development (Lai et al., 2006, 2008).

The most appropriate strategy for EST library development and sequencing depends on several factors, including the planned use of the library, whether a reference genome exists for the taxon being studied, and the financial resources available (Bouck and Vision, 2007; Mardis, 2008; Wall et al., 2009; Wheat, 2010). In this study, the main purpose of EST sequencing was gene discovery in the targeted weeds. As a consequence, in many instances, we isolated RNA from multiple tissue types and normalized the libraries to increase the likelihood of sampling rare transcripts (Tables 2, 3). Also, because sequencing technology has changed dramatically over the past decade, we have employed several different sequencing platforms, which vary in read length, types and rates of sequencing error, and cost per base pair (Mardis, 2008; Suzuki et al., 2011). Therefore, we can assess the value of increased tissue sampling and normalization relative to sequence depth for gene discovery, as well as potential trade-offs between sequencing depth and read length in the development of de novo transcriptome assemblies.

As commonly reported for other systems (Ohlrogge and Benning, 2000; Heesacker et al., 2008; Wall et al., 2009), sequencing from multiple tissue types and from normalized libraries did enhance gene discovery in the weeds targeted by this study

TABLE 4. Comparison of de novo assemblies of Illumina sequence data.

Taxon	Collection ID	Assembler	No. unigenes	Total assembly length (Mbp)	% UCOs [a]	% Full-length transcripts [b]	% Paralogs with Ks < 0.1
Cirsium arvense	Hodgins KN-ON	CLC	54 718	30	97	16.9	1.1
		Trinity	60 610	35	98	21.7	56.3
	Guggisberg et al., 280808-2	CLC	46 807	25	98	22.00	1.1
		Trinity	65 276	46	96	30	54.1
Helianthus annuus	SAW-3	CLC	37 108	13	83	13.7	0.9
		Trinity	45 804	20	87	21.4	65.8
Lactuca serriola	US96UC23	CLC	66 733	61	100	43.8	1.0
		Trinity	68 204	62	100	44.7	47.5

[a] Percentage of ultra-conserved orthologs (UCOs) found in EST library.

[b] Percentage of full-length transcripts calculated for UCOs.

TABLE 5. Microarrays developed for Compositae weeds targeted by this study.

Taxon	Platform	Collection ID	No. unigenes	No. features
Ambrosia artemisiifolia	12-plex	HU1-11	45 063	134 996
Centaurea diffusa	12-plex	DK TR001-1L	61 024	136 906
Centaurea solstitialis	4-plex	Multiple genotypes	34 343	68 686
Cirsium arvense	12-plex	NW-22-1-M	63 690	136 582
Helianthus annuus	4-plex	Several cultivars	33 376	68 400
Helianthus annuus	12-plex	Hann ISI	48 683	136 454

(Tables 2, 3). The advantage of sequencing from multiple tissues, however, appears to be surprisingly modest. For example, the percentage of ultra-conserved orthologs increased from 83% and 85% in libraries of common ragweed that were developed from leaf tissue to 91% in the GNV8ASA01 library from giant ragweed, which was sequenced to approximately the same depth, but included RNA from four tissues (Table 3). Normalization had a larger effect, with an increase in the fraction of ultra-conserved orthologs detected from 56% in a standard library of dandelion to 80% and 79% for normalized libraries of similar depth for spotted knapweed and yellow starthistle (Ceso JH1 #1), respectively (Table 3).

As expected, greater sequencing depth (Table 3) was correlated with detection of a higher fraction of ultra-conserved orthologs (Pearson's $r = 0.59$; df = 20; $P = 0.002$). Likewise, sequencing depth was strongly correlated with total assembly length for the Sanger and 454 libraries ($r = 0.80$; df = 17; $P = 0.000$), but this correlation was somewhat weaker when the Illumina data were included ($r = 0.53$; df = 20; $P = 0.008$), presumably because of variation in the quality of the de novo assemblies of Illumina data (see below).

While next-generation sequencing platforms provide a low cost method for obtaining large quantities of transcriptome data for nonmodel organisms, concerns have been expressed about the quality of de novo assemblies deriving from these platforms (Kumar and Blaxter, 2010; Robertson et al., 2010; Surget-Groba and Montoya-Burgos, 2010), especially the failure to distinguish between close paralogs (Barker et al., 2010) and to assemble full-length transcripts (Grabherr et al., 2011). While paralog discrimination may not be a major issue for molecular biologists, it is critical for population genomic studies and evolutionary analyses, such as the detection of whole genome duplications (Barker et al., 2008). As a consequence, 454 sequencing, which generates read lengths of 300–500 bp, has often been employed for the development of reference transcriptomes for nonmodel organisms (Vera et al., 2008; Peng et al., 2010; Prentis et al., 2010), despite its much greater expense when compared to the Illumina or ABI SOLiD platforms.

Our initial assembly results were generally consistent with these earlier observations. Assemblies of Sanger and 454 reads successfully distinguished between close paralogs, as measured by the proportion of duplicate genes with Ks < 0.1 (range = 22–67%, mean = 44%; Table 3). In contrast, our Illumina mRNA-Seq assemblies with CLC failed to resolve close paralogs, with the percentage of duplicates with Ks < 0.1 averaging 1.0% (Table 3). However, CLC and many other short read assemblers were developed for whole genome assemblies and are not optimal for the assembly of transcriptomes, which are expected to include huge variation in transcript coverage, as well as multiple kinds of transcripts per locus due to alternative splicing. Trinity, a recently published assembler program designed specifically for transcriptome data, is claimed to solve many of these issues (Grabherr et al., 2011). Our preliminary assemblies of Illumina transcriptome data indicate that close paralogs are resolved as claimed and that the program is more effective than aggressive assemblers such as CLC at recovering full-length transcripts (Table 4). Thus, it might be that the longer reads generated by Sanger or 454 are no longer necessary to generate reference transcriptomes.

Detection of hybridization—We tested several pairs of taxa for significant evidence of hybridization and introgression: common ragweed–giant ragweed, diffuse knapweed–spotted knapweed, common sunflower–cultivated sunflower, and prickly lettuce–cultivated lettuce. For the two ragweed species, and for wild sunflower, EST libraries were available for multiple accessions, which allowed us to test whether genomic patterns of hybridization and introgression, if it occurred, were similar across multiple contact zones.

For the ragweed and lettuce comparisons, only a single peak was observed in the Ks range examined, corresponding to the divergence between the two taxa. The two ragweed species showed a single, broad peak centered at Ks = 0.033+/−0.02 (Fig. 1A), regardless of populations compared, while the two lettuce species had a single peak at Ks = 0.08+/−0.005. This is the pattern expected if there has been no hybridization or introgression. The lack of evidence of introgression in the giant

TABLE 6. Accession numbers for EST libraries reported in this study.

Taxon	Collection ID	GenBank accessions or doi [a]
Ambrosia artemisiifolia	HU1-11	doi:10.5061/dryad.cm7td/12
	AA8-20	doi:10.5061/dryad.cm7td/3
Ambrosia trifida	GNV8ASA01	doi:10.5061/dryad.cm7td/7
	GNV8ASA02	doi:10.5061/dryad.cm7td/8
	GNV8ASA03	doi:10.5061/dryad.cm7td/9
	GNV8ASA04	doi:10.5061/dryad.cm7td/10
Carthamus oxyacanthus	PI 407602	doi:10.5061/dryad.cm7td/15
Centaurea diffusa	DK TR001-IL	doi:10.5061/dryad.cm7td/5
	DK US022-31E	doi:10.5061/dryad.cm7td/6
Centaurea stoebe subsp. *micranthos*	Cema #1A	GI:124612349-124626419
Centaurea solstitialis	Ceso JH1 #1	GI:124655902-124696404
	AR-13-24	doi:10.5061/dryad.cm7td/4
Cirsium arvense	NW-22-1-M	doi:10.5061/dryad.cm7td/14
	Hodgins KN-ON	doi:10.5061/dryad.cm7td/13
	Guggisberg et al., 280808-2	doi:10.5061/dryad.cm7td/2
Helianthus annuus	Hann ISI	doi:10.5061/dryad.cm7td/11
	SAW-3	doi:10.5061/dryad.cm7td/1
	Several cultivars [a]	See below [b]
Helianthus ciliaris	Hcil 1411	GI:125400397-125421999
Lactuca serriola	US96UC23 (Sanger library)	GI:83901317-83921492; 22397573-22415583; 22430445-22449769
	US96UC23 (Illumina library)	JO020427 - JO087153
Taraxacum officinale	Taof A68	GI:90246684- 90345856

[a] Transcriptome assemblies uploaded to Dryad data repository, available at http://dx.doi.org/10.5061/dryad.cm7td; doi = digital object identifier.

[b] Sanger EST libraries for *H. annuus* previously described by Heesacker et al. (2008), available at website http://cgpdb.ucdavis.edu/asteraceae_assembly/data_sequence_files/GB_ESTs_Feb_2007.sp.Heli_annu.clean.fasta.

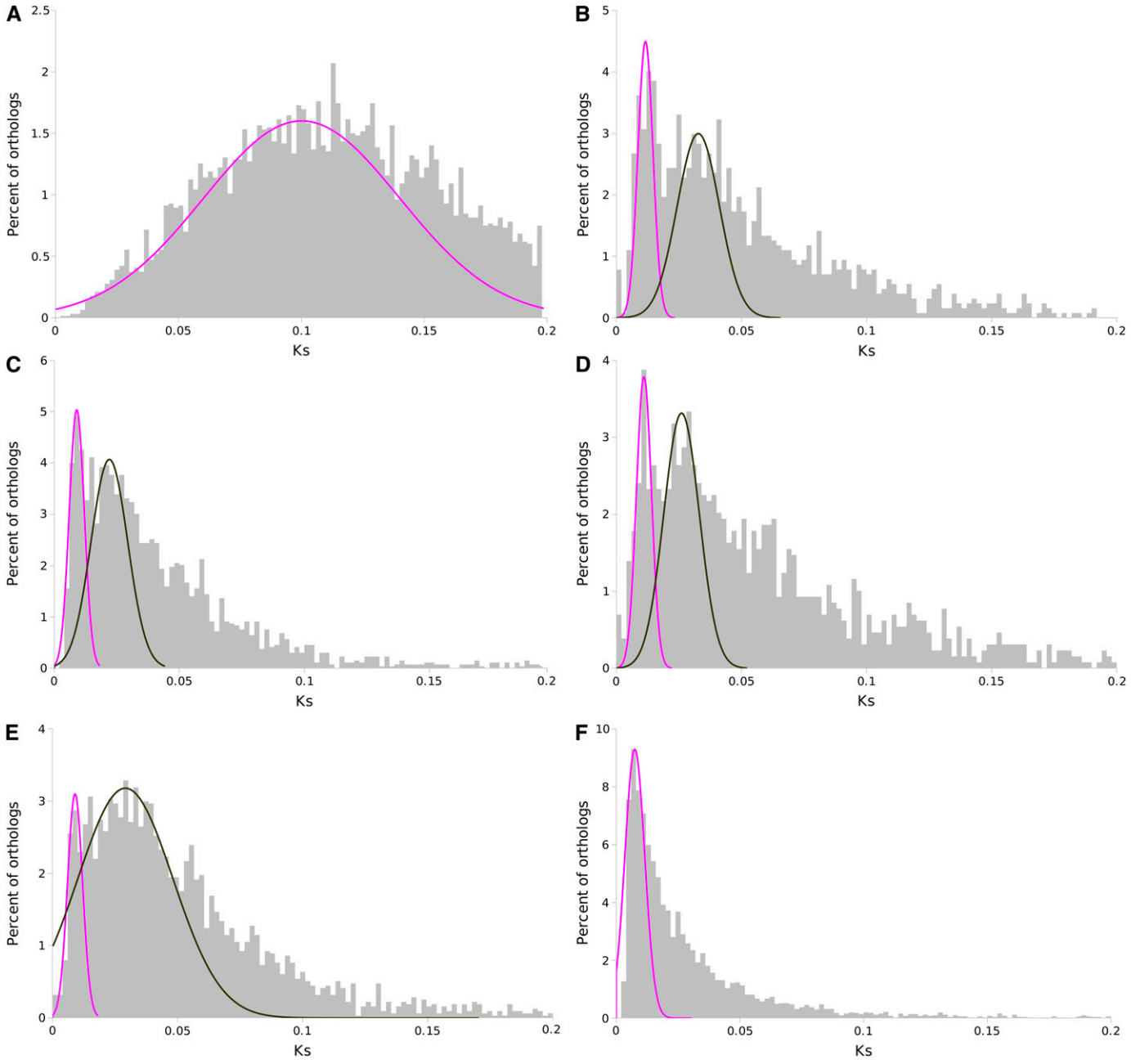

Fig. 1. Ks distributions and fitted normal curves for all ortholog pairs from the EMMIX analysis for four representative taxa: (A) Common vs. giant ragweed (AA8-20 vs. GNV8ASA01). (B) Diffuse knapweed from invasive range vs. spotted knapweed (DKUS022-31E vs. Cema #1A). (C) Weedy vs. domesticated sunflower (SAW3 vs. cultivars). (D) Diffuse knapweed from native range vs. spotted knapweed (DKTR001-1L vs. Cema #1A). (E) Weedy vs. domesticated sunflower (Hann - ISI vs. cultivars). (F) Prickly vs. domesticated lettuce (US96UC23 vs. cultivars).

ragweed populations was surprising, since we identified three plants in one of the invasive populations (GNV8ASA01) that were intermediate in morphology and genome size between common and giant ragweed (Q. Yu, unpublished data). However, pollen tube growth rates of hybrid pollen are greatly reduced in this cross (Vincent and Cappadocia, 1987). Thus, hybrid pollen is likely to be outcompeted by parental pollen, perhaps accounting for the apparent lack of backcrossing and introgression between the two species in nature.

In contrast, both the knapweed and sunflower comparisons showed two strongly significant peaks, indicating that introgression has altered the Ks distribution from that expected for divergence without gene flow. Diffuse knapweed from the invasive range had peaks at Ks = 0.012+/−0.003 and Ks = 0.033+/−0.008 (Fig. 1B), comprising 24% and 37%, respectively, of all ortholog pairs, while the native sample had peaks at Ks = 0.010+/−0.003 and Ks = 0.026+/−0.007 (Fig. 1D), comprising 20% and 26%, respectively, of all ortholog pairs. Note that the first peak in each comparison likely results from introgression, whereas the second corresponds to the divergence of the two species. Thus, the extent of introgression appears to be greater in the diffuse knapweed from the invasive than native

range, as previously reported by Blair and Hufbauer (2010) based on AFLP data. However, the introgression reported here likely occurred between diploid genotypes of the two species prior to their invasion of North America. In North America, the two species differ in ploidy, which appears to limit ongoing introgression (Blair and Hufbauer, 2010). Thus, highly introgressed genotypes of diffuse knapweed appear to have colonized North America.

The GO analysis showed significant differences in the function of genes in the introgression peaks of the knapweed samples, especially in the invasive sample. In the invasive comparison, proteins involved in development are overrepresented; proteins targeted to chloroplast or "unknown" are underrepresented; those targeted to ER, extracellular processes, and ribosome are overrepresented; proteins with functions as hydrolase or transferase are underrepresented; and transcription factors, receptors, other membrane proteins, or protein kinases are overrepresented. In the native range, the significant differences are limited to "other cellular components" and are much less significant, probably because the introgression peak in the native range comparison is smaller and less well defined.

Common sunflower from Australia had signs of introgression from domesticated sunflower, with peaks at Ks = 0.009+/−0.003 as well as Ks = 0.021+/−0.007 (Fig. 1C), comprising 18% and 29%, respectively, of all ortholog pairs. Weedy sunflower from Israel showed a less pronounced but still significant peak at 0.011+/−0.003 as well as 0.026+/−0.006, comprising 14% and 36%, respectively, of all ortholog pairs. These results are consistent with reports based on analyses of microsatellites that weedy sunflowers from outside North America may have a crop–wild ancestry (Muller et al., 2011). No significant biases in introgression patterns were detected by GO analyses in either of the sunflower comparisons, which is consistent with the lack of reproductive barriers between wild and domesticated populations of common sunflower.

An important caveat in the interpretation of these results is that they are based on comparisons between individual genotypes, which may not be representative of the population or taxon as a whole. Also, Ks comparisons between individual genotypes are noisy because they do not account for variation due to the coalescent or evolutionary rate heterogeneity. Thus, future analyses would be stronger if a population approach were taken, but this approach has been cost prohibitive until very recently. Nonetheless, our results demonstrate the power of this approach for detecting hybridization and introgression and for studying the kinds of genes that are most likely to introgress. By analyzing thousands of genes, robust conclusions can be made from noisy data.

Tools for analyses of gene expression and regulation—One of the main motivations for the EST sequencing reported here was to generate reference transcriptomes for Compositae weeds that could be used for studies of gene expression and regulation. Such studies are underway for five Compositae weeds (common ragweed, Canada thistle, yellow starthistle, diffuse knapweed, and common sunflower) and exploit the reference transcriptomes (Tables 2–4) and NimblegGen microarrays (Table 5) reported here. The 12-plex NimbleGen expression arrays represent an especially cost-effective strategy for population studies of expression variation, since only a handful of arrays are required for comprehensive analyses of gene expression patterns. Nonetheless, with the reduction in sequencing prices, it has become more cost feasible to study expression by deep sequencing of the transcriptome (Simon et al., 2009) in nonmodel organisms. However, even sequence-based studies of gene expression will require a reference transcriptome (or fully sequenced genome) for analyses, so the resources reported here will continue to be useful for expression studies of Compositae weeds.

Conclusions—We have generated EST resources and microarrays for 11 Compositae weeds, which we hope will facilitate studies of the origin and evolution of Compositae weeds, as well as the molecular basis of weedy traits in this group such as herbicide resistance (e.g., Peng et al., 2010) or growth-defense trade-offs (e.g., Mayrose et al., 2011). The resources presented were developed over 11 years, mainly by the Compositae Genome Project (http://compgenomics.ucdavis.edu/), and thus also demonstrate how strategies have been continuously refined to exploit advances in high-throughput sequencing and computational biology. Most recently, the development of the Trinity de novo assembler of short-read transcriptome data may allow reference-quality transcriptomes to be developed from very low-cost Illumina or ABI SOLiD sequence data (Grabherr et al., 2011), which could greatly reduce the cost of entry for genomic studies of nonmodel organisms.

Our study also demonstrates the utility of ortholog comparisons for identifying hybridization and quantifying the extent of introgression (Kane et al., 2009). While several authors have discussed the apparent association between hybridization and invasiveness (Abbott, 1992; Ellstrand and Schierenbeck, 2000; Rieseberg et al., 2007), it generally is not clear whether hybridization is a cause or consequence of range expansions (although see Whitney et al., 2006, 2010). Analyses of genomic data provide a sensitive and robust approach for detecting hybridization and introgression and for investigating whether introgressed variants have contributed to adaptive changes in weeds and invasive plants.

LITERATURE CITED

Abbott, R. J. 1992. Plant invasions, interspecific hybridization and the evolution of new plant taxa. *Trends in Ecology & Evolution* 7: 401–405.

Albert, V. A., D. E. Soltis, J. E. Carlson, W. G. Farmerie, P. K. Wall, D. C. Ilut, T. M. Solow, et al. 2005. Floral gene resources from basal angiosperms for comparative genomics research. *BMC Plant Biology* 5: 5.

Alexander, J. M. 2010. Genetic differences in the elevational limits of native and introduced *Lactuca serriola* populations. *Journal of Biogeography* 37: 1951–1961.

Arias, D. M., and L. H. Rieseberg. 1994. Gene flow betwen cultivated and wild sunflowers. *Theoretical and Applied Genetics* 89: 655–660.

Baack, E. J., Y. Sapir, M. A. Chapman, J. M. Burke, and L. H. Rieseberg. 2008. Selection on domestication traits and quantitative trait loci in crop–wild sunflower hybrids. *Molecular Ecology* 17: 666–677.

Barker, M. S., K. M. Dlugosch, L. Dinh, R. S. Challa, N. C. Kane, M. G. King, and L. H. Rieseberg. 2010. EvoPipes.net: Bioinformatic tools for ecological and evolutionary genomics. *Evolutionary Bioinformatics* 6: 143–149.

Barker, M. S., N. C. Kane, M. Matvienko, A. Kozik, W. Michelmore, S. J. Knapp, and L. H. Rieseberg. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Molecular Biology and Evolution* 25: 2445–2455.

Basu, C., M. D. Halfhill, T. C. Mueller, and C. N. Stewart. 2004. Weed genomics: New tools to understand weed biology. *Trends in Plant Science* 9: 391–398.

BELDADE, P., S. RUDD, J. D. GRUBER, AND A. D. LONG. 2006. A wing expressed sequence tag resource for *Bicyclus anynana* butterflies, an evo-devo model. *BMC Genomics* 7: 130.

BLAIR, A. C., AND R. A. HUFBAUER. 2010. Hybridization and invasion: One of North America's most devastating invasive plants shows evidence for a history of interspecific hybridization. *Evolutionary Applications* 3: 40–51.

BOUCK, A., AND T. VISION. 2007. The molecular ecologist's guide to expressed sequence tags. *Molecular Ecology* 16: 907–924.

BURKE, J. M., AND L. H. RIESEBERG. 2003. Fitness effects of transgenic disease resistance in sunflowers. *Science* 300: 1250.

CHAO, W. S., D. P. HORVATH, J. V. ANDERSON, AND M. E. FOLEY. 2005. Potential model weeds to study genomics, ecology, and physiology in the 21st century. *Weed Science* 53: 929–937.

CHAUDHURI, P., AND J. S. MARRON. 1999. SiZer for exploration of structures in curves. *Journal of the American Statistical Association* 94: 807–823.

CHAUVEL, B., F. DESSAINT, C. CARDINAL-LEGRAND, AND F. BRETAGNOLLE. 2006. The historical spread of *Ambrosia artemisiifolia* L. in France from herbarium records. *Journal of Biogeography* 33: 665–673.

CHEVREUX, B., T. PFISTERER, B. DRESCHER, A. J. DRIESEL, W. E. G. MULLER, T. WETTER, AND S. SUHAI. 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research* 14: 1147–1159.

COLAUTTI, R. I., S. A. BAILEY, C. D. A. VAN OVERDIJK, K. AMUNDSEN, AND H. J. MACISAAC. 2006. Characterised and projected costs of nonindigenous species in Canada. *Biological Invasions* 8: 45–59.

DAISIE. 2009. Handbook of alien species in Europe. Springer, Dordrecht, Netherlands.

DECHAINE, J. M., J. C. BURGER, AND J. M. BURKE. 2010. Ecological patterns and genetic analysis of post-dispersal seed predation in sunflower (*Helianthus annuus*) crop–wild hybrids. *Molecular Ecology* 19: 3477–3488.

DECHAINE, J. M., J. C. BURGER, M. A. CHAPMAN, G. J. SEILER, R. BRUNICK, S. J. KNAPP, AND J. M. BURKE. 2009. Fitness effects and genetic architecture of plant–herbivore interactions in sunflower crop-wild hybrids. *New Phytologist* 184: 828–841.

DEMPEWOLF, H., L. H. RIESEBERG, AND Q. C. CRONK. 2008. Crop domestication in the Compositae: A family-wide trait assessment. *Genetic Resources and Crop Evolution* 55: 1141–1157.

DEXTRASE, A. J., AND N. E. MANDRAK. 2006. Impacts of alien invasive species on freshwater fauna at risk in Canada. *Biological Invasions* 8: 13–24.

ELLIS, J. R., AND J. M. BURKE. 2007. EST-SSRs as a resource for population genetic analyses. *Heredity* 99: 125–132.

ELLSTRAND, N. C. 2003. Dangerous liaisons? When cultivated plants mate with their wild relatives. Johns Hopkins, Baltimore, Maryland, USA.

ELLSTRAND, N. C., AND K. A. SCHIERENBECK. 2000. Hybridization as a stimulus for the evolution of invasiveness in plants? *Proceedings of the National Academy of Sciences, USA* 97: 7043–7050.

FORMAN, J. 2003. The introduction of American plant species into Europe. *In* L. Child, J. Brock, and G. Brundu [eds.], Plant invasions: Ecological threats and management solutions, 17–39. Backhuys, Leiden, Netherlands.

GENTON, B. J., J. A. SHYKOFF, AND T. GIRAUD. 2005. High genetic diversity in French invasive populations of common ragweed, *Ambrosia artemisiifolia*, as a result of multiple sources of introduction. *Molecular Ecology* 14: 4275–4285.

GRABHERR, M. G., B. J. HAAS, M. YASSOUR, J. Z. LEVIN, D. A. THOMPSON, I. AMIT, X. ADICONIS, ET AL. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 9: 644–652

HEESACKER, A., V. K. KISHORE, W. X. GAO, S. X. TANG, J. M. KOLKMAN, A. GINGLE, M. MATVIENKO, ET AL. 2008. SSRs and INDELs mined from the sunflower EST database: Abundance, polymorphisms, and cross-taxa utility. *Theoretical and Applied Genetics* 117: 1021–1029.

HEISER, C. 2003. Weeds in my garden: Observations on some misunderstood plants. Timber Press, Portland, Oregon, USA.

HEISER, C., D. SMITH, S. CLEVENGER, AND W. MARTIN. 1969. The North American sunflowers (*Helianthus*). *Memoirs of the Torrey Botanical Club* 22: 1–218.

HUANG, X. Q., AND A. MADAN. 1999. CAP3: A DNA sequence assembly program. *Genome Research* 9: 868–877.

KANE, N. C., M. G. KING, M. S. BARKER, A. RADUSKI, S. KARRENBERG, Y. YATABE, S. J. KNAPP, ET AL. 2009. Comparative genomics and population genetic analyses indicate highly porous genomes and high leves of gene flow between divergent *Helianthus* species. *Evolution* 63: 2061–2075.

KOZIK, A., M. MATVIENKO, I. KOZIK, H. VAN LEEUWEN, A. VAN DEYNZE, AND R. M. MICHELMORE. 2008. Eukaryotic ultra conserved orthologs and estimation of gene capture in EST libraries. Plant and Animal Genome XVI Conference, P6, San Diego, California, USA.

KUMAR, S., AND M. L. BLAXTER. 2010. Comparing *de novo* assemblers for 454 transcriptome data. *BMC Genomics* 11: 571.

LAAIDI, M., K. LAAIDI, J. P. BESANCENOT, AND M. THIBAUDON. 2003. Ragweed in France: An invasive plant and its allergenic pollen. *Annals of Allergy, Asthma & Immunology* 91: 195–201.

LAI, Z., B. L. GROSS, Y. ZOU, J. ANDREWS, AND L. H. RIESEBERG. 2006. Microarray analysis reveals differential gene expression in hybrid sunflower species. *Molecular Ecology* 15: 1213–1227.

LAI, Z., N. C. KANE, Y. ZOU, AND L. H. RIESEBERG. 2008. Natural variation in gene expression between wild and weedy populations of *Helianthus annuus. Genetics* 179: 1881–1890.

LAI, Z., K. LIVINGSTONE, Y. ZOU, S. A. CHURCH, S. J. KNAPP, J. ANDREWS, AND L. H. RIESEBERG. 2005. Identification and mapping of SNPs from ESTs in sunflower. *Theoretical and Applied Genetics* 111: 1532–1544.

LAI, Z., Y. ZOU, N. C. KANE, J. H. CHOI, X. WANG, AND L. H. RIESEBERG. 2011. Preparation of normalized cDNA libraries for 454 Titanium transcriptome sequencing. *In* Population genomics: Methods and protocols, in press. Humana Press, New York, New York, USA.

LALONDE, R. G., AND B. D. ROITBERG. 1994. Mating system, life-history, and reproduction in Canada thistle (*Cirsium arvense* Asteraceae). *American Journal of Botany* 81: 21–28.

LAMB, C. E., P. H. RATNER, C. E. JOHNSON, A. J. AMBEGAONKAR, A. V. JOSHI, D. DAY, N. SAMPSON, ET AL. 2006. Economic impact of workplace productivity losses due to allergic rhinitis compared with select medical conditions in the United States from an employer perspective. *Current Medical Research and Opinion* 22: 1203–1210.

LEBEDA, A., I. DOLEZALOVA, V. FERAKOVA, AND D. ASTLEY. 2004. Geographical distribution of wild *Lactuca* species (Asteraceae, Lactuceae). *Botanical Review* 70: 328–356.

LEJEUNE, K. D., AND T. R. SEASTEDT. 2001. *Centaurea* species: The forb that won the west. *Conservation Biology* 15: 1568–1574.

LINDER, C. R., I. TAHA, G. J. SEILER, A. A. SNOW, AND L. H. RIESEBERG. 1998. Long-term introgression of crop genes into wild sunflower populations. *Theoretical and Applied Genetics* 96: 339–347.

MARDIS, E. R. 2008. Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics* 9: 387–402.

MAYROSE, M., N. KANE, I. MAYROSE, AND L. H. RIESEBERG. 2011. Increased vigour in sunflower correlates with reduced defenses and altered gene expression during biotic and abiotic stress responses. *Molecular Ecology*: doi:10.1111/j.1365-294X.2011.05301.x

MCLACHLAN, G., D. PEEL, K. BASFORD, AND P. ADAMS. 1999. The EMMIX software for the fitting of mixtures of normal and *t*-components. *Journal of Statistical Software* 4: 1-4.

MEYER, E., G. V. AGLYAMOVA, S. WANG, J. BUCHANAN-CARTER, D. ABREGO, J. K. COLBOURNE, B. L. WILLIS, ET AL. 2009. Sequencing and *de novo* analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics* 10: 219.

MULLER, M. H., M. LATREILLE, AND C. TOLLON. 2011. The origin and evolution of a recent agricultural weed: Population genetic diversity of weedy populations of sunflower (*Helianthus annuus* L.) in Spain and France. *Evolutionary Applications* 4: 499–514.

MUTH, N. Z., AND M. PIGLIUCCI. 2006. Traits of invasives reconsidered: Phenotypic comparisons of introduced invasive and introduced non-invasive plant species within two closely related clades. *American Journal of Botany* 93: 188–196.

MYERS, J., AND D. BAZELY. 2003. Ecology and control of introduced plants. Cambridge University Press, Cambridge, UK.

OHLROGGE, J., AND C. BENNING. 2000. Unraveling plant metabolism by EST analysis. *Current Opinion in Plant Biology* 3: 224–228.

PENG, Y. H., L. L. G. ABERCROMBIE, J. S. YUAN, C. W. RIGGINS, R. D. SAMMONS, P. J. TRANEL, AND C. N. STEWART. 2010. Characterization of the horseweed (*Conyza canadensis*) transcriptome using GS-FLX 454 pyrosequencing and its application for expression analysis of candidate non-target herbicide resistance genes. *Pest Management Science* 66: 1053–1062.

PIMENTEL, D., L. LACH, R. ZUNIGA, AND D. MORRISON. 2000. Environmental and economic costs of nonindigenous species in the United States. *BioScience* 50: 53–65.

PIMENTEL, D., R. ZUNIGA, AND D. MORRISON. 2005. Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecological Economics* 52: 273–288. doi:10.1016/j.ecolecon.2004.10.002

PRENTIS, P. J., M. WOOLFIT, S. R. THOMAS-HALL, D. ORTIZ-BARRIENTOS, A. PAVASOVIC, A. J. LOWE, AND P. M. SCHENK. 2010. Massively parallel sequencing and analysis of expressed sequence tags in a successful invasive plant. *Annals of Botany* 106: 1009–1017.

R DEVELOPMENT CORE TEAM. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Website http://www.R-project.org.

RICE, P. M. 2011. INVADERS Database System [online]. Website http://invader.dbs.umt.edu [accessed May 2011]. Division of Biological Sciences, University of Montana, Missoula, Montana, USA.

RIESEBERG, L. H., S. C. KIM, R. A. RANDELL, K. D. WHITNEY, B. L. GROSS, C. LEXER, AND K. CLAY. 2007. Hybridization and the colonization of novel habitats by annual sunflowers. *Genetica* 129: 149–165.

ROBERTSON, G., J. SCHEIN, R. CHIU, R. CORBETT, M. FIELD, S. D. JACKMAN, K. MUNGALL, ET AL. 2010. *De novo* assembly and analysis of RNA-seq data. *Nature Methods* 7: 909–912.

SIMBERLOFF, D. 2005. Non-native species do threaten the natural environment! *Journal of Agricultural & Environmental Ethics* 18: 595–607.

SIMON, S. A., J. X. ZHAI, R. S. NANDETY, K. P. MCCORMICK, J. ZENG, D. MEJIA, AND B. C. MEYERS. 2009. Short-read sequencing technologies for transcriptional analyses. *Annual Review of Plant Biology* 60: 305–333.

SNOW, A. A., D. PILSON, L. H. RIESEBERG, M. J. PAULSEN, N. PLESKAC, M. R. REAGON, D. E. WOLF, ET AL. 2003. A Bt transgene reduces herbivory and enhances fecundity in wild sunflowers. *Ecological Applications* 13: 279–286.

STERKY, F., R. R. BHALERAO, P. UNNEBERG, B. SEGERMAN, P. NILSSON, A. M. BRUNNER, L. CHARBONNEL-CAMPAA, ET AL. 2004. A *Populus* EST resource for plant functional genomics. *Proceedings of the National Academy of Sciences, USA* 101: 13951–13956.

STEVENS, P. 2001 [onward]. Angiosperm Phylogeny Website, version 9, June 2008 [and more or less continuously updated]. Website http://www.mobot.org/MOBOT/research/APweb/ [accessed 22 June 2011].

STEWART, C. N., P. J. TRANEL, D. P. HORVATH, J. V. ANDERSON, L. H. RIESEBERG, J. H. WESTWOOD, C. A. MALLORY-SMITH, ET AL. 2009. Evolution of weediness and invasiveness: Charting the course for weed genomics. *Weed Science* 57: 451–462.

STORMS, W., E. O. MELTZER, R. A. NATHAN, AND J. C. SELNER. 1997. The economic impact of allergic rhinitis. *Journal of Allergy and Clinical Immunology* 99: S820–S824.

SURGET-GROBA, Y., AND J. I. MONTOYA-BURGOS. 2010. Optimization of *de novo* transcriptome assembly from next-generation sequencing data. *Genome Research* 20: 1432–1440.

SUZUKI, S., N. ONO, C. FURUSAWA, B. W. YING, AND T. YOMO. 2011. Comparison of sequence reads obtained from three next-generation sequencing platforms. *PLoS ONE* 6: e19534. *doi/10.1371/journal.pone.0019534*

VERA, J. C., C. W. WHEAT, H. W. FESCEMYER, M. J. FRILANDER, D. L. CRAWFORD, I. HANSKI, AND J. H. MARDEN. 2008. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology* 17: 1636–1647.

VERDUIJN, M. H., P. J. VAN DIJK, AND J. M. M. VAN DAMME. 2004. The role of tetraploids in the sexual-asexual cycle in dandelions (*Taraxacum*). *Heredity* 93: 390–398.

VINCENT, G., AND M. CAPPADOCIA. 1987. Interspecific hybridization between common ragweed (*Ambrosia artemisiifolia*) and giant ragweed (*A. trifida*). *Weed Science* 35: 633–636.

WALL, P. K., J. LEEBENS-MACK, A. S. CHANDERBALI, A. BARAKAT, E. WOLCOTT, H. Y. LIANG, L. LANDHERR, ET AL. 2009. Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics* 10: 347. doi:10.1186/1471-2164-10-347.

WANG, Y., AND J. HEY. 2010. Estimating divergence parameters with small samples from a large number of loci. *Genetics* 184: 363–379.

WHEAT, C. W. 2010. Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. *Genetica* 138: 433–451.

WHITNEY, K. D., R. A. RANDELL, AND L. H. RIESEBERG. 2006. Adaptive introgression of herbivore resistance traits in the weedy sunflower *Helianthus annuus. American Naturalist* 167: 794–807.

WHITNEY, K. D., R. A. RANDELL, AND L. H. RIESEBERG. 2010. Adaptive introgression of abiotic tolerance traits in the sunflower *Helianthus annuus. New Phytologist* 187: 230–239.

WILCOVE, D. S., D. ROTHSTEIN, J. DUBOW, A. PHILLIPS, AND E. LOSOS. 1998. Quantifying threats to imperiled species in the United States. *Bioscience* 48: 607–615.