MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2013) doi: 10.1111/1755-0998.12163

Genomics of Compositae crops: reference transcriptome assemblies and evidence of hybridization with wild relatives

KATHRYN A. HODGINS,* ZHAO LAI,† LUIZ O. OLIVEIRA,‡ DAVID W. STILL,§ MOIRA SCASCITELLI,* MICHAEL S. BARKER,¶ NOLAN C. KANE,** HANNES DEMPEWOLF,* ALEX KOZIK,†† RICHARD V. KESSELI,‡‡ JOHN M. BURKE,§§ RICHARD W. MICHELMORE,††¶¶ and LOREN H. RIESEBERG*†

*Department of Botany and Biodiversity Research Centre, University of British Columbia, Vancouver, BC V6T 1Z4, Canada, †Department of Biology and Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN 47405, USA, ‡Departamento de Bioquímica e Biologia Molecular, Universidade Federal de Viçosa, 36570-000 Viçosa, Brazil, §Department of Plant Sciences, Cal Poly Pomona, Pomona CA 91768, USA, ¶Department of Ecology and Evolutionary Biology, University of Arizona, Tucson AZ 85721, USA, **Department of Ecology and Evolutionary Biology, University of Colorado Boulder, Boulder, CO 80309, USA, ††The Genome Center, University of California, Davis, CA 95616, USA, ‡‡Department of Biology, University of Massachusetts, Boston, MA 02125, USA, §§Department of Plant Biology, University of Georgia, Athens, GA 30602, USA, ¶¶Departments of Plant Sciences, Molecular & Cellular Biology, and Medical Microbiology & Immunology, University of California, Davis, CA 95616, USA

Abstract

Although the Compositae harbours only two major food crops, sunflower and lettuce, many other species in this family are utilized by humans and have experienced various levels of domestication. Here, we have used next-generation sequencing technology to develop 15 reference transcriptome assemblies for Compositae crops or their wild relatives. These data allow us to gain insight into the evolutionary and genomic consequences of plant domestication. Specifically, we performed Illumina sequencing of *Cichorium endivia*, *Cichorium intybus*, *Echinacea angustifolia*, *Iva annua*, *Helianthus tuberosus*, *Dahlia hybrida*, *Leontodon taraxacoides* and *Glebionis segetum*, as well 454 sequencing of *Guizotia scabra*, *Stevia rebaudiana*, *Parthenium argentatum* and *Smallanthus sonchifolius*. Illumina reads were assembled using Trinity, and 454 reads were assembled using MIRA and CAP3. We evaluated the coverage of the transcriptomes using BLASTX analysis of a set of ultra-conserved orthologs (UCOs) and recovered most of these genes (88–98%). We found a correlation between contig length and read length for the 454 assemblies, and greater contig lengths for the 454 compared with the Illumina assemblies. This suggests that longer reads can aid in the assembly of more complete transcripts. Finally, we compared the divergence of orthologs at synonymous sites (*Ks*) between Compositae crops and their wild relatives and found greater divergence when the progenitors were self-incompatible. We also found greater divergence between pairs of taxa that had some evidence of postzygotic isolation. For several more distantly related congeners, such as chicory and endive, we identified a signature of introgression in the distribution of *Ks* values.

Keywords: Compositae, crop genomics, hybridization, introgression, transcriptome

Received 22 May 2013; revision received 14 August 2013; accepted 15 August 2013

Introduction

The domestication of plants represented a critical development in human history that permitted the establishment of large, sedentary civilizations. Consequently, unravelling the origin of crops, as well as the molecular and genetic changes that accompany domestication and crop diversification represents an important undertaking. Until recently, such studies were confined to a handful of crops of major economic and nutritional importance, such as maize, wheat and rice (Burger *et al.* 2008). However, advances in next-generation sequencing technology have allowed the extension of genomic knowledge beyond these species to a wider array of crops and their wild relatives (e.g. Dempewolf *et al.* 2010; Agarwal *et al.* 2012; Scaglione *et al.* 2012). This information will not only be an important agronomic resource, but it will also improve the understanding of the genomic basis of domestication and adaptation.

Correspondence: Kathryn A. Hodgins, Fax: 604-822-6089; E-mail: hodgins@zoology.ubc.ca

The Compositae (Asteraceae) is one of the largest and most successful flowering plant families. Despite the large number of species in this family, only two – sunflower and lettuce – have become major food crops. However, there are many other species in the Compositae that have been cultivated by humans and attained various degrees of domestication. Although the number of species in the Compositae that have been strongly domesticated is disproportionately small compared with some other groups, such as the Fabaceae or Poaceae, no other family has been cultivated for such a wide variety of uses (Dempewolf *et al.* 2008). Species in the family have been domesticated for seed oil (e.g. sunflower), edible leaves (e.g. lettuce), edible inflorescences or stems (e.g. globe artichoke), tubers and roots (e.g. yacon), phytochemicals (e.g. guayule), and ornamental flowers (e.g. gerbera). This diversity of uses makes investigations into the genomic basis of domestication in this group particularly interesting.

Here, we describe the development of genomic resources for 12 Compositae species (Table 1): *Cichorium endivia* (chicory, wild and cultivated), *Cichorium intybus* (endive, wild and cultivated), *Echinacea angustifolia*, *Iva annua* (sumpweed), *Helianthus tuberosus* (Jerusalem artichoke), *Dahlia hybrida*, *Leontodon taraxacoides*, *Glebionis segetum* (corn chrysanthemum), *Guizotia scabra* ssp. *schimperii*, *Stevia rebaudiana* (sweetleaf), *Parthenium argentatum* (guayule) and *Smallanthus sonchifolius* (yacón). Most of these species are crops, or crop wild relatives, and have been cultivated for a wide variety of uses. Similar to lettuce, chicory and endive are native to the Old World and are grown mainly for their edible leaves, although chicory is also grown for its tubers (Kiers *et al.* 2000). Echinacea, native to North America, is cultivated

Table 1 Location information for Compositae crops and their wild relatives targeted in this study and the tissue type sampled

Taxon	Common name	Collection locality	Collection ID	Tissue type
Cichorium endivia ssp. *pumilum* (wild)	Endive	Pakistan	PI 652029	Seedling
Cichorium endivia ssp. *endivia* (cultivar)	Endive	Germany	PI 503595	Seedling
Cichorium intybus (wild)	Chicory	Krasnodar, Russian Federation (latitude 45.033, longitude 35.977)	PI 652028	Seedling
Cichorium intybus (cultivar)	Chicory – Witloof	Germany	PI 504468	Seedling
Dahlia hybrida	Dahlia 'Thomas Edison'	NA	NA	Leaves, flowers
Echinacea angustifolia (wild)	Coneflower	Oklahoma, United States, Section 24, T19N, R2W, Logan County 36.1–97.367	PI 421331	Seedling
Glebionis segetum	Corn chrysanthemum	Cleden-cap-Sizun, Finistere, France	PI 586603	Seedling
Glebionis segetum (wild)	Corn chrysanthemum	Porto, Portugal. Between Lordelo do Ouro and Porto, Douro Litoral Province. Latitude 41.15, Longitude -8.633	PI 641689	Seedling
Iva annua	Sumpweed	Granite City, IL Latitude 38.804 Longitude -90.114	NA	Seedling
Leontodon taraxacoides Lam. ssp. *saxatilis*	Lesser hawkbit	Oregon, Benton City, OSU campus, vacant lot at corner of SW 11th St and Washington	NA	Seedling
Helianthus tuberosus (wild)	Jerusalem artichoke	Ohio, United States, Hwy. 81W, 16.8 km west of Ada, Allen County. Latitude 40.733, Longitude -84.017	PI 547230	Seedling
Guizotia scabra (wild) ssp. *schimperii*	Mech	Jimma. Ethiopia. 5 km from Jimma on the way to Bonga, 1775 m evolution. Latitude 7.626, Longitude 36.760	RC-4	Seedling
Stevia rebaudiana	Sweetleaf	Garden origin, West Coast Seeds, B.C. Canada	NA	Seedling
Parthenium argentatum	Guayule	NA	PI 478640	Roots, leaves, flowers, stem
Smallanthus sonchifolius	Yacon	Peru	CIP 205029	Roots, leaves, flowers, stem

for its believed immunostimulator properties (Percival 2000). There are several oil producing seed crops in the Compositae including sunflower, safflower (*Carthamus tinctorius*) and noug (*Guizotia abyssinica*), of which *G. scabra* ssp. *schimperii* is thought to be its closest living wild relative. Sumpweed was once cultivated by North American First Nations people for its edible seeds, but was abandoned prior to the arrival of Europeans, perhaps due to its allergenic properties (Diamond 1997). Yacón and Jerusalem artichoke, both of which are New World crops, have been domesticated for their inulin-rich tuberous roots (Dempewolf *et al.* 2008). Many species of the Compositae are cultivated as ornamentals, including dahlias, originating mainly in Mexico (Saar *et al.* 2003), and chrysanthemums. Corn chrysanthemum, native to Europe and the Mediterranean, is grown as an ornamental, but is no longer considered to be part of the economically important florist chrysanthemum genus (Paciolla *et al.* 2010). Sweetleaf, native to Paraguay, is propagated as a sweetener (Brandle *et al.* 1998), and guayule, native to the south-west United States and Mexico, is cultivated as a source of natural rubber (Ray 1993). We have also sequenced the transcriptome of *Leontodon taraxacoides*, a weed originating in Europe and introduced into the United States, which is being developed as a small genome model for the Compositae. The genome size of this species (0.29 Gb, E. Baack & L. H. Rieseberg, unpublished) is dwarfed by other members of the family, which usually have genomes exceeding 1 Gb (Bennett & Leitch 2012).

The Compositae also harbours many of the world's most notorious weeds and several Compositae crops are closely related to weedy taxa. Genomic resources will be valuable for detecting gene flow between various crops and their wild and weedy relatives. Such gene flow can have implications for the spread of genetically engineered genes from crops into wild species (Ellstrand 2003; Snow *et al.* 2003) or contamination of seed lots by foreign germplasm (Bateman 1947a,b; Warburton *et al.* 2011). More generally, the study of gene flow between domesticated species and their progenitors could give insight into the strength of reproductive barriers and the process of speciation (Dempewolf *et al.* 2012), as well as the evolutionary consequences of hybridization and introgression (Hufford *et al.* 2013). Although there have been an increasing number of studies using genetic markers to estimate gene flow between cultivated and weedy populations (Arias & Rieseberg 1994; Ellstrand 2003; Song *et al.* 2003; Hufford *et al.* 2013), there have been few genome-wide studies especially across multiple crop/wild species pairs.

Two Compositae crops, lettuce and sunflower are particularly interesting with respect to their histories of domestication and invasiveness. Sunflower was domesticated in North America, yet today, *H. annuus* and many taxa in the genus are naturalized or invasive in Europe (Rehorek 1997; Forman 2003). High levels of gene flow between cultivated and weedy sunflower are known to occur (Arias & Rieseberg 1994; Linder *et al.* 1998; Burke *et al.* 2002) fuelling debates about transgene escape in the evolution of 'super weeds' (e.g. Burke & Rieseberg 2003; Ellstrand 2003; Snow *et al.* 2003). Lettuce was domesticated in the Mediterranean region, yet its progenitor, prickly lettuce (*L. serriola*), and several other taxa in the genus are considered weeds in the United States. There are also concerns about wild-crop gene flow in a number of other Compositae crops, including chicory (Kiær *et al.* 2007, 2009) and safflower (Berville *et al.* 2005a). These concerns are often well founded due to sympatry of crops with weedy and wild relatives, high outcrossing rates, and few postzygotic barriers between crops and their progenitors or feral weeds.

The genomic resources that we have developed are a valuable resource for future population and comparative genomic analyses of crop and weed evolution. We also compared *de novo* assemblies across platforms and correlated features of the read sets and the final assemblies to examine the impact of different sequencing strategies on the quality of the final assemblies. In addition, we used the resulting assemblies as well as others previously generated by the Compositae Genome Project, to examine divergence between crops and their putative progenitors and to consider evidence for introgression between crops and their wild relatives.

Materials and methods

Library development and sequencing

Upon harvesting tissue (see Table 1 for tissue types), we instantly froze it in liquid nitrogen and then stored it in a −80°C freezer. We extracted total RNA using the TRIzol reagent (Invitrogen)/RNeasy (QIAGEN) approach as described in Lai *et al.* (2006). For 454 sequencing (454 Life Sciences, Branford, CT, 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. For yacón, 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. Then, normalized cDNA was prepared for sequencing following the standard genomic DNA shotgun protocol recommended by 454 Life 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, we fragmented the cDNA to 500-to-800-bp fragments by either sonication or nebulization and size selected to remove small fragments using AMPure SPRI beads (Angencourt, Beverly, MA, 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.* 2012b).

For Illumina sequencing, we used two different methods. For dahlia, a non-normalized RNA-seq library was prepared by the Genome Sciences Center as recommended by Illumina. For the remaining species, cDNA was synthesized using the SMART PCR cDNA Synthesis Kit (Clontech, Palo Alto, CA, USA) and then normalized with the TRIMMER-DIRECT Kit. We then prepared the normalized libraries for sequencing as recommended by Illumina. After, we determined the fragment size distributions on a Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) and the concentrations with PicoGreen (Invitrogen).

We sequenced the 454 EST libraries on GS-FLX machines (454 Life Sciences) at the David H. Murdock Research Institute (DHMRI) or Genome Quebec using the standard 454 Titanium chemistry. We sequenced dahlia at the GSC on the Illumina GAII, and the remaining Illumina libraries on the UBC BioDiversity Research Center Illumina HiSeq2000.

Transcriptome processing and assembly

We used the pipeline SnoWhite version 2.0.2 (http://evopipes.net/snowhite.html; Barker *et al.* 2010) to filter the Illumina and 454 data using a minimum phred score set to 20. Illumina data were assembled with the program Trinity (http://trinityrnaseq.sourceforge.net/) using the butterfly parameters – 'bfly_opts, edge-thr=0.05' – to increase its ability to distinguish close paralogs. We used MIRA version 3.2.1 (Chevreux *et al.* 2004) using the 'accurate.est.denovo.454' assembly mode to assemble the 454 data. As MIRA can split up high coverage contigs, we used CAP3 at 94% identity to further assemble the MIRA contigs and singletons (Huang & Maden 1999). We only retained contigs from the MIRA and CAP3 assemblies >200 bp to make the Trinity assemblies and the MIRA/CAP3 assemblies comparable.

We assessed the assembly quality using a number of metrics. We examined the number of unigenes, the assembly length, average unigene length, maximum unigene length, as well as the proportion of ultra-conserved orthologs (UCOs) detected using the NCBI program BLASTX and an *e*-value threshold of 10^{-10} and an alignment length of 30 amino acids or more. The UCOs are 357 single-copy genes that are shared by *Arabidopsis thaliana*, humans, mice, yeast, fruit flies and *Caenorhabditis elegans* (A. Kozik, unpublished; http://compgenomics.ucdavis.edu/compositae_reference.php). We also determined the percentage of UCOs that were at least 80% of the length of the corresponding *A. thaliana* protein to estimate the number of transcripts that were close to full length in our assembly. Lastly, we evaluated the proportion of recently duplicated paralogs in the assembly, by analyzing duplicate gene age distributions using the DupPipe pipeline (Barker *et al.* 2008, 2010). We performed this analysis because assemblies of short reads or overaggressive assemblies can fail to distinguish between recently diverged paralogs. We used linear models to compare platforms and examine how features of the read set influence metrics of assembly quality (lm in R).

Ortholog identification and divergence

For several Compositae crops, we had transcriptomes from the crop and their wild relatives [globe artichoke (*Cynara cardunculus var. scolymus*), cardoon (*Cynara cardunculus* var. *altilis*)], Jerusalem artichoke, noug, safflower, chicory, endive, lettuce and sunflower (Table S2, Supporting information). Using assemblies available from the Compositae Genome Project website (Barker *et al.* 2008; Heesacker *et al.* 2008; Lai *et al.* 2012a; Scaglione *et al.* 2012; Dempewolf *et al.* 2008) as well as those from this publications, we identified orthologs between congeneric taxa using reciprocal best hits and determined the number of synonymous substitutions per synonymous site (*Ks*) using the same methods as described in Kane *et al.* (2009). Briefly, using results from a BLASTX to the UniProt plant protein database, we predicted proteins using hidden Markov models with the program Wise 2.2 (Birney *et al.* 2004). Following this, we aligned orthologous proteins with MUSCLE3.7 (Edgar 2004), reverse translated the alignments with RevTrans1.4 (Wernersson & Pedersen 2003) and used codeml from PAML4.5 to calculate *Ks* values (Yang 1997, 2007). We examined the average *Ks* values between crops and their wild relatives to determine if features of their domestication history, or reproductive barriers were associated with the degree of divergence. Specifically, we examined whether the strength of the domestication syndrome (Dempewolf *et al.* 2008), mating system, self-incompatibility system (crop and progenitor) and history of domestication (Table S4, Supporting information) were associated with divergence. We also used reports in the literature to determine whether there were any postzygotic barriers to gene flow, such as reduced crossability of the crop and the wild relative, reduced fitness of the F1 hybrids or later generations (Table S4, Supporting information).

Many of these crop taxa co-occur and potentially hybridize with their wild relatives. To test for hybridization, we examined the distribution of *Ks* values for orthologs, which should be centred around a *Ks* value reflecting the time as the most recent common ancestor of the taxa involved, while a secondary peak at a lower *Ks* value can be attributed to more recent gene flow (Wang & Hey 2010). To identify significant peaks in the ortholog *Ks* distribution, we used SiZer (Chaudhuri & Marron 1999). Then, we used EMMIX (McLachlan *et al.* 1999) to determine the location of significant peaks in the range $0 < Ks < 0.1$. The optimal number of peaks was inferred as the model that minimizes the Bayesian information criterion. 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 between putative introgressed genes and nonintrogressed genes using Fisher's exact test in R (R Core Team, 2013) and used the false discovery rate comparison to correct for multiple tests (Storey 2002).

Databases

We have archived the assemblies generated by this study on the Compositae Genome Project Database (http://compgenomics.ucdavis.edu/) and the raw data at the Short Read Archive (SRP020001). In addition, we have placed the reference assemblies in Dryad (doi:10.5061/dryad.cp723).

Results

Transcriptome sequencing

We sequenced, on average, 3.1 Gb using the Illumina platform representing 34.4 Gb of sequence in total for 10 libraries in eight different species (Table S1, Supporting information). Our stringent filtering with the pipeline SnoWhite resulted in a large reduction of reads for the Illumina assemblies (mean read reduction = 23.2%). This was largely due to the removal of short reads after trimming adapters, poly-A/T, and low-quality ends. For the four 454 assemblies, we sequenced on average 479.13 Mb for four different species. The filtering of the 454 data resulted in only 1.9% decrease on average in the number of reads for *G. scabra* ssp. *schimperii* and sweetleaf, but yacón and guayule had a larger number of reads removed due to the shorter length of the original reads for these libraries (Table S1, Supporting information).

The assembly statistics for the Illumina and 454 transcriptomes are shown in Table 2. The number of unigenes, as well as the total length of the assembly, was significantly correlated with the total amount of sequence data for the Illumina assemblies (contig number: $t_9 = 2.42$, $P < 0.05$; assembly length: $t_9 = 4.62$, $P < 0.01$), but not the 454 assemblies (contig number: $t_2 = 0.23$, $P = 0.84$; assembly length: $t_2 = 0.71$, $P = 0.55$), probably partly due to the small number of 454 assemblies. For the Trinity assemblies, the number of components (unique isoforms) was also correlated with the amount of sequence, although the relationship was not as strong ($t_9 = 2.12$, $P = 0.063$). For the 454 assemblies, average length of the contigs and the number of long UCO transcripts were associated with read length, despite the small sample size (contig length: $t_2 = 6.35$, $P < 0.05$; UCO length: $t_2 = 6.63$, $P < 0.05$). Read length was not tested for the Illumina platform as all the libraries had the same read length (100 bp paired end) with the exception of dahlia. The total amount of sequence was not correlated with average contig length for the Illumina data, but it was marginally significant for the 454 platform (Illumina: $t_9 = -0.28$, $P = 0.78$; 454: $t_2 = 3.79$, $P < 0.10$). The percentage of close paralogs (% paralogs with $Ks < 0.1$) was not correlated with the amount of sequence data for either platform (Illumina: $t_9 = 1.09$, $P = 0.31$; 454: $t_2 = 0.90$, $P = 0.46$), although it was negatively correlated with the average length of the unigenes for Illumina, but not 454 (Illumina: $t_9 = -3.38$, $P < 0.01$; 454: $t_2 = 0.52$, $P = 0.65$). Similarly, the percentage of close paralogs was positively correlated with the total number of unigenes for Illumina, but not 454 (Illumina: $t_9 = 3.64$, $P < 0.01$; 454: $t_2 = -1.59$, $P = 0.25$).

The average length of the unigenes differed between the two platforms (Illumina mean = 585; 454 mean = 828; $t_{13} = -4.29$, $P < 0.001$), and the total assembly length was marginally significant (Illumina mean = 42.14 Mbp; 454 mean = 54.3 Mbp; $t_{13} = -1.95$, $P = 0.077$). The two platforms did not differ significantly in the number of unigenes (Illumina mean = 74 326; 454 mean = 64 983; $t_{13} = -1.53$, $P = 0.15$) and the percentage of long UCOs ($t_{13} = -0.078$, $P = 0.93$; Illumina mean = 43.4; 454 mean = 43.9). The number of UCOs recovered was high and was over 94% for all of the assemblies except for one *G. segetum* Illumina assembly and the *P. argentatum* 454 assembly. These two assemblies had among the lowest numbers of Mbp sequenced for their respective platforms. For the Illumina assemblies, the number of reads was relatively low for *de novo* assemblies after filtering (mean = 20.1 million reads), but we were still able to

Table 2 Assembly statistics for the Compositae crop transcriptomes targeted in this study

Taxon	Collection ID	Sequence type	Read No.	Total sequence (Mbp)	Contig No.	Total assembly length (Mbp)	Average length	Max unigene length	% UCOs	% Full length	% Paralogs with Ks < 0.1
Cichorium endivia ssp. *pumilum* (wild)	CHE-24285	Illumina	181 24 638	2157.07	52 685	34.02	646	8355	96.9	49.9	27.8
Cichorium endivia ssp. *endivia* (cultivar)	CHE-3178	Illumina	37 847 448	4504.40	63 647	46.04	723	8684	98.0	59.7	26.2
Cichorium intybus (wild)	CHI-2418	Illumina	24 881 864	2961.28	56 696	38.78	684	4690	96.6	49.0	32.9
Cichorium intybus (cultivar)	CHI-Witloof	Illumina	22 213 574	2643.70	58 926	37.40	635	4583	95.5	46.8	32.6
Dahlia hybrida (ornamental)	Dahlia	Illumina	41 908 902	4886.02	135 229	64.15	474	7324	98.9	39.2	41.3
Echinacea angustifolia (wild)	Echinacea	Illumina	23 122 980	2751.93	89 514	43.84	490	5403	98.6	35.3	38.8
Glebionis segetum	GLEB-21149	Illumina	27 900 884	3320.53	82 554	44.46	539	5322	95.0	40.1	38.6
Glebionis segetum (wild)	GLEB-25383	Illumina	26 030 484	3097.98	57 996	30.51	526	3961	89.4	23.3	33.1
Iva annua	Iva	Illumina	20 261 030	2411.36	67 250	36.72	547	3090	96.4	42.6	27.1
Leontodon taraxacoides	Leontodon	Illumina	22 344 074	2659.23	63 338	41.03	648	4805	98.6	53.2	28.1
Helianthus tuberosus (wild)	TUB-2047	Illumina	25 005 770	2976.02	89 749	46.61	519	4192	98.9	38.4	37.0
Guizotia scabra ssp. *schimperii* (wild)	RC4	454	1 267 082	608.65	63 189	58.7	930	8990	98.0	56.9	57.4
Stevia rebaudiana	Sta	454	1 252 897	604.20	60 028	54.3	905	8020	94.1	56.0	58.4
Parthenium argentatum	Guayule	454	983 076	266.48	51 947	33.4	642	4943	88.0	27.2	48.9
Smallanthus sonchifolius	Yacon	454	1 324 268	437.19	84 767	70.8	835	8271	95.0	35.6	34.6

detect most of the UCOs (mean = 96.6%). Similarly, for the 454 assemblies, we had significant blast hits to most of the UCOs (mean = 93.8%). The ability to detect close paralogs differed significantly between the platforms (Illumina mean = 33.02 ± 1.6; 454 mean = 49.81 ± 5.51; $t_{13} = -4.08$, $P < 0.001$).

Divergence of crops and wild relatives and evidence of hybridization

Using *de novo* transcriptome assemblies from the Compositae Genome Project (Barker *et al.* 2008; Heesacker *et al.* 2008; Dempewolf *et al.* 2010; Lai *et al.* 2012a; Scaglione *et al.* 2012), including those published here, we were able to compare divergence of orthologs at synonymous sites between crops and their wild relatives for nine species pairs (Tables 3, S2 and S4, Supporting information). For several species pairs (globe artichoke, cardoon, noug, lettuce and sunflower), we were able to compare multiple transcriptomes. The estimated average divergence agreed very well when replicated across multiple transcriptomes for the same species pairs, even when different sequencing platforms were used (Tables S2 and S3, Supporting information). For example, average divergence between wild cardoon and globe artichoke was identical for both 454 and Illumina

Table 3 The average synonymous substitution rates (*Ks*) for comparisons of Compositae crops and their relatives

Species comparison		*Ks*
Cynara cardunculus var. *scolymus*	*Cynara cardunculus* var. *sylvestris*	0.022
Cynara cardunculus var. *altilis*	*Cynara cardunculus* var. *sylvestris*	0.022
Helianthus tuberosus (cultivated)	*Helianthus tuberosus* (wild)	0.04
Guizotia abyssinica	*Guizotia abyssinica* ssp. *schimperi*	0.051
Carthamus tinctorius	*Carthamus palastinus*	0.027
Cichorium intybus (cultivated)	*Cichorium intybus* (wild)	0.034
Cichorium endivia ssp. *endivia*	*Cichorium endivia* ssp. *pumilum*	0.035
Lactuca sativa	*Lactuca serriola*	0.02
Helianthus annuus (cultivated)	*Helianthus annuus* (wild)	0.044
Cynara cardunculus var. *scolymus*	*Cynara cardunculus* var. *altilis*	0.022
Cichorium endivia	*Cichorium intybus*	0.065
Helianthus annuus	*Helianthus tuberosus*	0.053
Lactuca sativa	*Lactuca virosa*	0.057
Helianthus annuus (wild)	*Helianthus annuus* (weedy)	0.048
Carthamus tinctorius	*Carthamus oxyacanthus*	0.027

assemblies, and estimates of divergence between lettuce (*Lactuca sativa*) and prickly lettuce (*Lactuca virosa*) were very similar when either Sanger or Illumina prickly lettuce assemblies were used. The average divergence between crops and their wild relatives was 0.032 and ranged from 0.022 in cardoon and lettuce to 0.051 in noug. We also examined divergence between other closely related species that may have had a history of introgression (Tables 3; S2 and S3, Supporting information). Estimates of average divergence were the highest between endive and chicory (both cultivated and wild accessions) and were similar across all four comparisons (mean = 0.065). Divergence between sunflower (domesticated and wild accessions) and Jerusalem artichoke accessions was more variable and averaged 0.053 (range 0.049–0.065), while divergence between lettuce and bitter lettuce averaged 0.057.

We examined whether there were any features of the crop/progenitor pairs that might be associated with divergence at silent sites (Table S4, Supporting information). We found no association between divergence and the degree of domestication (semi-domesticated crops = 0.034 ± 0.007; strongly domesticated = 0.032 ± 0.004; $t_7 = -0.24$, $P = 0.83$), the history of domestication (<4000 years = 0.031 ± 0.004; >4000 years = 0.036 ± 0.007; $t_7 = -0.65$, $P = 0.54$), or self-compatibility in the crop (SC = 0.028 ± 0.004; SI = 0.042 ± 0.005; $t_7 = 2.06$, $P = 0.08$). However, self-incompatible progenitor species were significantly more divergent from their crop relatives compared with self-compatible progenitor species (SC = 0.025 ± 0.003; SI = 0.042 ± 0.004; $t_7 = 3.89$, $P < 0.01$). But, with the exception of chicory, which had the lowest divergence for this group, all of the self-incompatible progenitors were found in the Heliantheae or Millerieae tribes (Table S4, Supporting information). When the species in these tribes were compared to the other species, they had significantly higher average *Ks* values (other tribes = 0.027 ± 0.003; Heliantheae or Millerieae = 0.045 ± 0.003; $t_7 = 4.15$, $P < 0.01$). Only two crops/progenitor pairs (sunflower and noug) had some evidence of postzygotic isolation in the literature, although in both cases fertile offspring are still produced (Table S4, Supporting information). These two species had significantly greater average divergence from their progenitors compared with the other crop/progenitor pairs (no postzygotic barriers = 0.029 ± 0.003; postzygotic barriers = 0.048 ± 0.004; $t_7 = 3.17$, $P < 0.05$). When we included prezygotic barriers, such as assortative mating due to selfing, there was no difference in divergence between crops and their progenitors (no barriers = 0.028 ± 0.006; barriers = 0.035 ± 0.005; $t_7 = 0.92$, $P = 0.37$). A similar pattern was seen when we expanded the analysis to include comparisons of other congeneric species (e.g. sunflower vs Jerusalem artichoke; Table S4, Supporting information): species pairs with known postzygotic barriers had significantly higher divergence compared with those without known postzygotic barriers (no postzygotic barriers = 0.030 ± 0.003; postzygotic barriers = 0.054 ± 0.003; $t_{13} = 3.32$, $P < 0.01$), but when prezygotic barriers were also included, there was no significant difference (no barriers = 0.031 ± 0.006; barriers = 0.043 ± 0.005; $t_{13} = 1.35$, $P = 0.20$).

Most of the crops and their wild relatives exhibited little divergence across orthologs and the distribution of *Ks* values decayed exponentially, which is expected in comparisons of individuals from the same species, suggesting ongoing gene flow. Because of this, we were unable to examine the distribution of *Ks* values for multiple peaks to assess the presence of recent introgression for many crop–progenitor pairs. However, with noug and its progenitor, *G. scabra* ssp. *schimperii*, we found signs of divergence followed by recent gene flow (Fig. 1a; Table S3, Supporting information), with a recent peak at 0.0081 on average. No GO terms were significantly over-represented in the set of putatively introgressed genes for either of the two comparisons made. We also found evidence of introgression in the comparison of endive and chicory, which was replicated in all four comparisons of the wild and cultivated accessions, with a recent peak at 0.0085 on average (Figs 1b; S1, and Table S3, Supporting information). Similarly, we found evidence of gene flow in the comparison of lettuce and bitter lettuce (Fig. 1c), and in all three comparisons of sunflower (wild and domesticated accessions) and wild Jerusalem artichoke (Fig. 1d and S1, Supporting information). However, this pattern was not evident between sunflower (wild and domesticated accessions) and the cultivated Jerusalem artichoke accession. After FDR correction, one of the comparisons between chicory and endive had GO terms significantly over-represented (FDR rate = 5%) in the set of putatively introgressed genes. The biological processes that were over-represented were carbon fixation, ATP synthesis-coupled electron transport and mitotic cell cycle spindle assembly checkpoint (Table S3, Supporting information). No other GO terms were over-represented in our analysis using a FDR = 5%.

Discussion

Transcriptome sequencing is currently the most cost-effective method of gene discovery in nonmodel organisms. The central goal of this study was to identify gene sequences across a wide array of Compositae crops and their wild relatives, and to this end, we were able to assemble transcriptomes from 15 samples across 12 different species. These data add to the growing number of Compositae crop transcriptomes already available

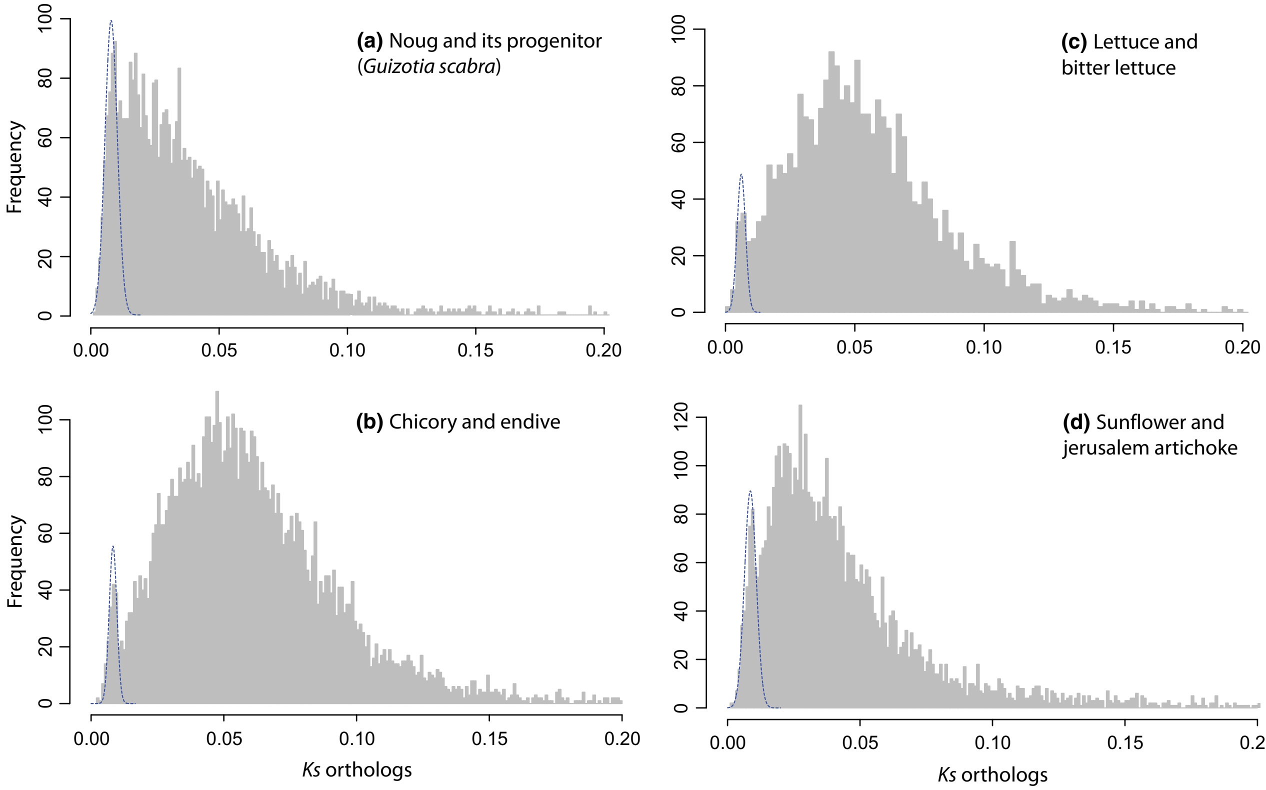

Fig. 1 *Ks* distributions for all ortholog pairs for four taxa. Fitted normal curves from the EMMIX analysis are shown for species where evidence for introgression was detected. (a) noug and its progenitor (*Guizotia scabra* ssp. *schimperii*). (b) chicory (PI 652028) and endive (PI 652029). (c) lettuce and bitter lettuce (*Lactuca virosa*). (d) sunflower (accession HA412) and Jerusalem artichoke (accession PI 547230).

and represent an important genomic resource for future studies.

Transcriptome sequencing methods

For *de novo* transcriptome assemblies where a broad characterization of the transcriptome is desired, several factors have been shown to impact gene discovery including sequencing depth, normalization and the life stages/tissue types sampled (Wall *et al.* 2009; Ekblom & Galindo 2011; Lai *et al.* 2012a; Matvienko *et al.* 2013). A predominant factor in determining the number of unigenes identified in our study was sequencing depth, as evidenced by the correlation between sequencing depth and unigene number, as well as assembly length for the Illumina assemblies. The number of tissues sampled also may have improved gene discovery. Most of the libraries that we sequenced were made from seedling tissue; however, for dahlia, yacón and guayule, we were able to sequence multiple tissue types (Table 1). Dahlia had the greatest number of unigenes, components (unique isoforms) and had the highest assembly length out of all the assemblies. Yacón had the largest number of unigenes and the longest assembly for the 454 data, despite having similar sequence depth to *G. scabra* ssp. *schimperii* and *S. rebaudiana,* perhaps owing to the multiple tissue types sampled. However, polyploidy could have also contributed to the high number of contigs in both cases (Gatt *et al.* 1998; Viehmannová *et al.* 2009). One method to increase the breadth of the transcriptome coverage is normalization of the cDNA library (Ekblom & Galindo 2011; Lai *et al.* 2012a). With the exception of dahlia, the use of normalized libraries in our study probably improved gene discovery. Consequently, we were able to identify most of the UCOs (>87% in all cases), suggesting that our sampling of the transcriptome was adequate to capture a large fraction of the genes across most species.

For the 454 assemblies, read length had a large impact on unigene length as both the average length of the unigenes and the number of long UCOs were correlated with read length. Also, the comparison of the average unigene length of the two platforms revealed a clear advantage to the 454 assemblies for this metric. Despite the high coverage, paired-end data and improved assembly algorithms (e.g. Grabherr *et al.* 2011; Schulz *et al.* 2012) available for short-read data, there are still benefits in obtaining the longer reads offered by platforms such as

Roche 454FLX or the Illumina MiSeq. Simulations have also found that a mixed platform approach can improve the number of full length genes through the inclusion of longer reads (Wall *et al.* 2009). Although higher short-read coverage could aid in recovering full-length genes (Wall *et al.* 2009), we found no correlation between average contig length and total sequence for the Illumina assemblies, even when the octoploid dahlia was removed from the analysis. For the Illumina data set, the percentage of close paralogs was negatively correlated with unigene length and positively correlated with unigene number. This suggests that the Trinity assembler, although much better at resolving paralogs and alternatively spliced transcripts than assemblers designed for genome assembly (Grabherr *et al.* 2011; Lai *et al.* 2012a), results in more fragmented assemblies when a larger number of paralogs were present. This could be due to a higher proportion of poorly represented isoforms in the transcript pool, as Trinity and other competing assemblers are known to reconstruct fewer full-length transcripts when transcript abundance is low (Grabherr *et al.* 2011; Schulz *et al.* 2012).

Divergence between crops and wild relatives

Domestication within the Compositae has been recent for most species, and reproductive barriers between crops and wild progenitors appear to be weak or absent in many cases, suggesting that they are the same biological species (Table S4, Supporting information). Thus, it is perhaps not surprising that we failed to find any association between divergence at synonymous sites and time since domestication or the strength of the domestication syndrome. Average divergence of synonymous sites from samples taken from the same species should reflect the species' effective population size, *Ne* (Wang & Hey 2010). We found that self-incompatible progenitors had greater average divergence from their corresponding crops compared with self-compatible progenitors. *Ne* is predicted to be lower within self-fertilizing species as a consequence of high homozygosity, life histories that promote genetic bottlenecks and reduced interspecific gene flow (Charlesworth & Wright 2001; Wright *et al.* 2008). Indeed, the impact of mating system on within-population neutral diversity has been frequently demonstrated (Schoen & Brown 1991; Glemin *et al.* 2006). Also, many of the species that were self-incompatible were from the tribes Heliantheae or Millerieae. The exception was chicory, which had the lowest divergence of the self-incompatible species. When Heliantheae and Millerieae species were compared to the other species, these tribes had significantly higher average *Ks* values. There is evidence of a shared whole-genome duplication at the base of these tribes (Barker *et al.* 2008) and perhaps erroneous inclusion of some paralogs inflated the divergence estimates for these comparisons.

Reproductive barriers should result in the accumulation of genetic differences between populations (Dobzhansky 1940; Coyne & Orr 1989; Hendry & Taylor 2004). We found that the two species with the greatest divergence between the crop and the putative progenitor (noug and sunflower) had evidence for postzygotic isolating barriers, although these barriers are weak in the case of sunflower and the high average *Ks* value probably reflects the large effective population size of this species (Strasburg *et al.* 2011). However, when we included estimates of divergence and reproductive isolation of more distantly related congeners, this pattern was confirmed and is consistent with findings in many other plant and animal taxa (see Edmands 2002 for review).

Evidence for introgression

Most comparisons of crops and their closest wild and weedy relatives revealed little divergence, suggesting ongoing gene flow or recent divergence. In our comparisons of more distantly related taxa, we were able to identify evidence for introgression in a number of cases. We found clear evidence for introgression between chicory and endive in all comparisons. This finding is not unexpected as, although there is moderate sterility and reduced germination of the F1s, hybrids can be readily acquired under field conditions (Rick 1953). Similarly, noug and *G. scabra* ssp. *schimperi* showed evidence of divergence followed by gene flow. There is limited postzygotic isolation and relatively high crossability between noug and this taxon as well as others in the genus (Dagne 1994). Also, unlike sunflower, another oilseed crop in the family, noug does not exhibit strong signs of artificial selection and resembles its wild relatives to a greater degree, suggesting ongoing gene flow (Dempewolf 2011). However, population-level studies of noug and its putative progenitor using SSR markers have found little evidence of hybridization (Dempewolf 2011). The discrepancy between these results and ours could reflect differences in timescale, as SSR markers have much higher mutation rates compared with synonymous sites (Ossowski *et al.* 2010). Therefore, although noug and its progenitor have experienced gene flow in the past, it may not be ongoing. Alternatively, this difference could reflect differences in the history of hybridization among the individuals sampled for the two studies.

We found unexpected evidence for introgression in the case of lettuce and bitter lettuce as well as sunflower and Jerusalem artichoke. Although they are both diploid, extreme sterility of the hybrids results from crosses between lettuce and bitter lettuce (Hayes & Ryder 2007).

Bitter lettuce has been used occasionally in lettuce breeding efforts (Ryder 1979; Hayes & Ryder 2007). Also, the introgression identified in this analysis could reflect introgression from other species, or perhaps more ancient introgression, that occurred after the species diverged, but before reproductive isolation was as complete. Strong reproductive barriers are also found between sunflower and Jerusalem artichoke that reflect differences in ploidy. However, wild Jerusalem artichoke accessions have been used in breeding modern cultivated sunflower, so some of the signal may be due to gene flow into sunflower, but the fact that only one of Jerusalem artichoke accessions shows this peak indicates that much of the gene flow was introgression from sunflower into *H. tuberosus*. Insect-mediated hybridization between the two species has been observed, and annual sunflower has been used for improvement in some Jerusalem artichoke cultivars (Berville *et al*. 2005b). This may explain the evidence for introgression found in one of the comparisons between these two species.

Barriers to gene flow between crops and their wild relatives have important implications for transgene escape (Ellstrand 2003), as well as for the genetic resources available for breeding from wild germplasm (e.g. disease resistance, herbicide resistance, drought tolerance; Tanksley & McCouch 1997). The absence of many postzygotic isolating mechanisms, including ploidy differences, and several outcrossing or mixed mating progenitors/crops provide ample potential for transgene escape in many of the species considered here, but also abundant genetic variation potentially useful for future breeding efforts. Our divergence estimates reveal that there is likely ongoing or recent gene flow between wild populations and crops in several cases. Indeed several population studies using molecular markers to assess hybridization between wild and domesticated Compositae crops are consistent with this (Arias & Rieseberg 1994; Kiær *et al*. 2009; Uwimana *et al*. 2012). Similar to many weedy Composites, interspecific gene flow has left its imprint in the genome of several Compositae crops and their wild relatives. In the future, it will be interesting to determine whether any of these putatively introgressed genes have been important during domestication and crop improvement.

Acknowledgements

We thank the Kevin Kubeck, Peter Carrington, Eric Baack and the USDA for plant samples, and Megan Stewart, Harry Lutton, Marta Matvienko, Huaqin Xu and Jie Huang for technical assistance. This research was funded by the US National Science Foundation (DBI-0820451) and the Natural Sciences and Engineering Research Council of Canada (Grant 327475). LOO received fellowships from Coodenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES BEX 0586/08-0) and from Conselho Nacional de Desenvolvimento Científico e Tecnológico (PQ 304153/2012-5). We would also like to thank Ethiopian Institute of Biodiversity Conservation, Addis Ababa University and the Ethiopian Institute of Agricultural Research for their assistance, as well as DHMRI, Genome Quebec and the UBC BioDiversity Research Center for sequencing our samples.

References

Agarwal G, Jhanwar S, Priya P *et al.* (2012) Comparative analysis of kabuli chickpea transcriptome with desi and wild chickpea provides a rich resource for development of functional markers. *PLoS One*, **7**, e52443.

Arias D, Rieseberg LH (1994) Gene flow between cultivated and wild sunflowers. *Theoretical and Applied Genetics*, **89**, 655–660.

Barker MS, Kane NC, Matvienko M *et al.* (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.

Barker MS, Dlugosch KM, Dinh L *et al.* (2010) EvoPipes.net: bioinformatic tools for ecological and evolutionary genomics. *Evolutionary Bioinformatics*, **6**, 143–149.

Bateman AJ (1947a) Contamination of seed crops II. Wind pollination. *Heredity*, **1**, 235–246.

Bateman AJ (1947b) Contamination of seed crops: i insect pollination. *Journal of Genetics*, **48**, 257–275.

Beldade P, Rudd S, Gruber J, Long A (2006) A wing expressed sequence tag resource for *Bicyclus anynana* butterflies, an evo-devo model. *BMC Genomics*, **16**, 1–16.

Bennett MD, Leitch IJ (2012) Plant DNA *C*-values database (release 6.0, Dec. 2012) Available from http://www.kew.org/cvalues/.

Berville A, Breton C, Cunliffe K *et al.* (2005a) Issues of ferality or potential for ferality in oats, olives, the Vigna group, ryegrass species, safflower, and sugarcane. In: *Crop Ferality and Volunteerism* (ed Gressel J), pp. 231–255. CRC Press, Taylor & Francis Group, Boca Raton, Florida.

Berville A, Muller MH, Poinso B, Serieys H (2005b) Ferality risks of gene flow between sunflower and other *Helianthus*. In: *Crop Ferality and Volunteerism* (ed Gressel J), pp. 209–230. CRC Press, Taylor & Francis Group, Boca Raton, Florida.

Birney E, Clamp M, Durbin R (2004) Genewise and genomewise. *Genome Research*, **14**, 988–995.

Brandle JE, Starratt AN, Gijzen M (1998) *Stevia rebaudiana*: its agricultural, biological, and chemical properties. *Canadian Journal of Plant Science*, **78**, 527–536.

Burger JC, Chapman MA, Burke JM (2008) Molecular insights into the evolution of crop plants. *American Journal of Botany*, **95**, 113–122.

Burke JM, Rieseberg LH (2003) Fitness effects of transgenic disease resistance in sunflowers. *Science*, **300**, 1250.

Burke J, Gardner KA, Rieseberg LH (2002) The potential for gene flow between cultivated and wild sunflower (*Helianthus annuus*) in the United States. *American Journal of Botany*, **89**, 1550–1552.

Charlesworth D, Wright SI (2001) Breeding systems and genome evolution. *Current Opinion in Genetics & Development*, **11**, 685–690.

Chaudhuri P, Marron JS (1999) SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, **94**, 807–823.

Chevreux B, Pfisterer T, Drescher B *et al.* (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Research*, **14**, 1147–1159.

Coyne J, Orr H (1989) Patterns of speciation in *Drosophila*. *Evolution*, **43**, 362–381.

Dagne K (1994) Meiosis in interspecific hybrids and genomic interrelationships in Guizotia Cass. (Compositae). *Hereditas*, **121**, 119–129.

Dempewolf H (2011) *Patterns of domestication in the compositae and beyond*. PhD, University of British Columbia, Canada.

Dempewolf H, Rieseberg L, Cronk QC (2008) Crop domestication in the compositae: a family-wide trait assessment. *Genetic Resources and Crop Evolution*, **55**, 1141–1157.

Dempewolf H, Kane NC, Ostevik KL *et al.* (2010) Establishing genomic tools and resources for *Guizotia abyssinica* (L.f.) Cass. - the development of a library of expressed sequence tags, microsatellite loci, and the sequencing of its chloroplast genome. *Molecular Ecology Resources*, **10**, 1048–1058.

Dempewolf H, Hodgins KA, Rummell SE, Ellstrand NC, Rieseberg LH (2012) Reproductive isolation during domestication. *Plant Cell*, **24**, 2710–2717.

Diamond JM (1997) *Guns, Germs, and Steel: The Fates of Human Societies*. WW Norton & Co, New York.

Dobzhansky T (1940) Speciation as a stage in evolutionary divergence. *American Naturalist*, **74**, 312–321.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.

Edmands S (2002) Does parental divergence predict reproductive compatibility? *TRENDS in Ecology & Evolution*, **17**, 520–527.

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, **107**, 1–15.

Ellstrand NC (2003) *Dangerous Liaisons? When Cultivated Plants Mate with Their Wild Relatives*. Johns Hopkins University Press, Baltimore.

Forman J (2003) The introduction of American plant species into Europe: issues and consequences. In: *Plant Invasions: Ecological Threats and Management Solutions* (eds Child LE, Brock JH, Brundu G, Prach K, Pysek P, Wade PM & Williamson M), pp. 17–39. Backhuys Publishers, Leiden, The Netherlands.

Gatt M, Ding H, Hammett K, Murray B (1998) Polyploidy and evolution in wild and cultivated *Dahlia* species. *Annals of Botany*, **81**, 647–656.

Glemin S, Bazin E, Charlesworth D (2006) Impact of mating systems on patterns of sequence polymorphism in flowering plants. *Proceedings of the Royal Society Biological Sciences*, **273**, 3011–3019.

Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652.

Hayes RJ, Ryder EJ (2007) Introgression of novel alleles for partial resistance to big vein disease from *Lactuca virosa* into cultivated lettuce. *HortScience*, **42**, 35–39.

Heesacker A, Kishore VK, Gao W *et al.* (2008) Abundance, polymorphisms, and cross-taxa utility of sunflower EST-SSRs. *Theoretical and Applied Genetics*, **117**, 1021–1029.

Hendry AP, Taylor EB (2004) How much of the variation in adaptive divergence can be explained by gene flow? An evaluation using lake-stream stickleback pairs. *Evolution*, **58**, 2319–2331.

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

Hufford MB, Lubinksy P, Pyhäjärvi T *et al.* (2013) The genomic signature of crop-wild introgression in maize. *PLoS Genetics*, **9**, e1003477.

Kane NC, King MG, Barker MS *et al.* (2009) Comparative genomic and population genetic analyses indicate highly porous genomes and high levels of gene flow between divergent *Helianthus* species. *Evolution*, **63**, 1–15.

Kiær L, Philipp M, Jørgensen R, Hauser T (2007) Genealogy, morphology and fitness of spontaneous hybrids between wild and cultivated chicory (*Cichorium intybus*). *Heredity*, **99**, 112–120.

Kiær LP, Felber F, Flavell A *et al.* (2009) Spontaneous gene flow and population structure in wild and cultivated chicory, *Cichorium intybus* L. *Genetic Resources and Crop Evolution*, **56**, 405–419.

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

Lai Z, Gross BL, Zou Y, Andrews J, Rieseberg LH (2006) Microarray analysis reveals differential gene expression in hybrid sunflower species. *Molecular Ecology*, **15**, 1213–1227.

Lai Z, Kane NC, Kozik A *et al.* (2012a) Genomics of compositae weeds: EST libraries, microarrays, and evidence of introgression. *American Journal of Botany*, **99**, 1–10.

Lai Z, Zou Y, Kane NC, Choi JH, Wang X, Rieseberg LH (2012b) Preparation of normalized cDNA libraries for 454 Titanium transcriptome sequencing. *Methods in Molecular Biology*, **888**, 119–133.

Linder CR, Taha I, Seiler GJ, Snow AA, Rieseberg LH (1998) Long-term introgression of crop genes into wild sunflower populations. *TAG Theoretical and Applied Genetics*, **96**, 339–347.

Matvienko M, Kozik A, Froenicke L *et al.* (2013) Consequences of normalizing transcriptomic and genomic libraries of plant genomes using a duplex-specific nuclease and tetramethylammonium chloride. *PLoS One*, **8**, e55913.

McLachlan GJ, Peel D, Basford KE, Adams P (1999) The Emmix software for the fitting of mixtures of normal and t-components. *Journal of Statistical Software*, **4**, 1–4.

Meyer E, Aglyamova GV, Wang S *et al.* (2009) Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics*, **10**, 219.

Ossowski S, Schneeberger K, Lucas-Lledo JI *et al.* (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science*, **327**, 92–94.

Paciolla C, D'Emerico S, Tommasi F, Scrugli A (2010) Karyomorphological and biochemical studies in *Glebionis coronaria* (L.) Spach and *Glebionis segetum* (L.) Fourreau from Italy. *Plant Biosystems*, **144**, 563–567.

Percival SS (2000) Use of echinacea in medicine. *Biochemical Pharmacology*, **60**, 155–158.

R Core Team (2013) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org/

Ray DT (1993) Guayule: A source of natural rubber. In: *New Crops* (eds Simon JE & Janick J), pp. 338–343. Wiley, New York.

Rehorek V (1997) Cultivated and escaped perennial *Helianthus* species in Europe. *Preslia*, **69**, 59–70.

Rick C (1953) Chicory-endive hybridized: isolation necessary to prevent production of undesired hybrids by the two species. *California Agriculture*, **7**, 7.

Ryder E (1979) 'Salinas' lettuce. *HortScience*, **14**, 283–284.

Saar DE, Polans NO, Sorensen PD (2003) A phylogenetic analysis of the genus *Dahlia* (Asteraceae) based on internal and external transcribed spacer regions of nuclear ribosomal DNA. *Systematic Botany*, **28**, 627–639.

Scaglione D, Lanteri S, Acquadro A *et al.* (2012) Large-scale transcriptome characterization and mass discovery of SNPs in globe artichoke and its related taxa. *Plant Biotechnology Journal*, **10**, 956–969.

Schoen DJ, Brown AH (1991) Intraspecific variation in population gene diversity and effective population size correlates with the mating system in plants. *Proceedings of the National Academy of Sciences*, **88**, 4494–4497.

Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, **15**, 1086–1092.

Snow AA, Pilson D, Rieseberg LH *et al.* (2003) A Bt transgene reduces herbivory and enhances fecundity in wild sunflowers. *Ecological Applications*, **13**, 279–286.

Song ZP, Lu BR, Zhu YG, Chen JK (2003) Gene flow from cultivated rice to the wild species *Oryza rufipogon* under experimental field conditions. *New Phytologist*, **157**, 657–665.

Storey JD (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B*, **64**, 479–498.

Strasburg JL, Kane NC, Raduski AR, Bonin A, Michelmore R, Rieseberg LH (2011) Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. *Molecular Biology and Evolution*, **28**, 1569–1580.

Tanksley SD, McCouch SR (1997) Seed banks and molecular maps: unlocking genetic potential from the wild. *Science*, **277**, 1063–1066.

Uwimana B, D'Andrea L, Felber F *et al.* (2012) A Bayesian analysis of gene flow from crops to their wild relatives: cultivated (*Lactuca sativa* L.) and prickly lettuce (*L. serriola* L.) and the recent expansion of *L. serriola* in Europe. *Molecular Ecology*, **21**, 2640–2654.

Viehmannová I, Cusimamani EF, Bechyne M, Vyvadilová M, Greplová M (2009) *In vitro* induction of polyploidy in yacon (*Smallanthus sonchifolius*). *Plant Cell, Tissue and Organ Culture*, **97**, 21–25.

Wall PK, Leebens-Mack J, Chanderbali AS *et al*. (2009) Comparison of next generation sequencing technologies for transcriptome characterization. *BMC genomics*, **10**, 347–357.

Wang Y, Hey J (2010) Estimating divergence parameters with small samples from a large number of loci. *Genetics*, **184**, 363–379.

Warburton ML, Wilkes G, Taba S *et al.* (2011) Gene flow among different teosinte taxa and into the domesticated maize gene pool. *Genetic Resources Crop Evolution*, **58**, 1243–1261.

Wernersson R, Pedersen AG (2003) RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Research*, **31**, 3537–3539.

Wright SI, Ness RW, Foxe JP, Barrett SCH (2008) Genomic consequences of outcrossing and selfing in plants. *International Journal of Plant Sciences*, **169**, 105–118.

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences*, **13**, 555–556.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586–1591.

L.H.R., J.M.B., R.W.M. and Z.L. designed the research. Z.L., L.O. and M.S. performed the laboratory work. K.A.H., M.B. and N.C.K. performed the assemblies. A.K. developed the UCOs set. K.A.H. conducted the analysis and wrote the paper. D.W.S., R.V.K., L.H.R., Z.L., M.S. and H.D. provided tissue. All authors contributed to the final version of the study.

Data Accessibility

All the raw sequence data has been deposited in the NCBI SRA (SRP020001). In addition, the assemblies are available on the Compositae Genome Project website (http://compgenomics.ucdavis.edu/index.php) and in Dryad (doi:10.5061/dryad.cp723).

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1 The *Ks* distributions for pairwise comparisons of Compositae crops and their wild relatives.

Table S1 The read number and total amount of sequence data before and after filtering the reads.

Table S2 The assemblies used in the comparative analysis.

Table S3 The average divergence at synonymous sites between crops and their relatives.

Table S4 Domestication and life history traits of the Compositae crops and their wild relatives.