RESEARCH ARTICLE

AMERICAN JOURNAL OF BOTANY

BRIEF COMMUNICATION

Genome-wide genotyping-by-sequencing data provide a high-resolution view of wild *Helianthus* diversity, genetic structure, and interspecies gene flow[1]

Gregory J. Baute[2,*], Gregory L. Owens[2,*], Dan G. Bock[2], and Loren H. Rieseberg[2–4]

PREMISE: Wild sunflowers harbor considerable genetic diversity and are a major resource for improvement of the cultivated sunflower, *Helianthus annuus*. The *Helianthus* genus is also well known for its propensity for gene flow between taxa.

METHODS: We surveyed genomic diversity of 292 samples of wild *Helianthus* from 22 taxa that are cross-compatible with the cultivar using genotyping by sequencing. With these data, we derived a high-resolution phylogeny of the taxa, interrogated genome-wide levels of diversity, explored *H. annuus* population structure, and identified localized gene flow between *H. annuus* and its close relatives.

KEY RESULTS: Our phylogenomic analyses confirmed a number of previously established interspecific relationships and indicated for the first time that a newly described annual sunflower, *H. winteri*, is nested within *H. annuus*. Principal component analyses showed that *H. annuus* has geographic population structure with most notable subpopulations occurring in California and Texas. While gene flow was identified between *H. annuus* and *H. bolanderi* in California and between *H. annuus* and *H. argophyllus* in Texas, this genetic exchange does not appear to drive observed patterns of *H. annuus* population structure.

CONCLUSIONS: Wild *H. annuus* remains an excellent resource for cultivated sunflower breeding effort because of its diversity and the ease with which it can be crossed with cultivated *H. annuus.* Cases of interspecific gene flow such as those documented here also indicate wild *H. annuus* can act as a bridge to capture alleles from other wild taxa; continued breeding efforts with it may therefore reap the largest rewards.

KEY WORDS ABBA BABA; Asteraceae; crop wild relatives; GBS; *Helianthus annuus*; *Helianthus* hybridization; phylogenetics; sunflower

Capitalizing on the biodiversity maintained in the world's seed banks is expected to be a critical component of obtaining and maintaining food security into the future (McCouch et al., 2013). Historically, wild progenitors and closely related congeners of modern crops have been an important source of high-value traits in crop improvement programs (Hajjar and Hodgkin, 2007). Genotype data on wild germplasm can facilitate breeding by allowing diversity to be catalogued and maximized among a limited number of wild donors used. Furthermore, gene flow among wild species can allow potentially beneficial alleles to travel from a difficult donor species to one easily bred into domestic germplasm (Jansky and Hamernik, 2009). Genotypic data of multiple species allows us to identify where this gene flow may be occurring. Low cost high-throughput sequencing is making genomic surveys of crop wild relatives (CWRs) possible even for crops with modest genomic resources and can provide a basis for classifying accessions, establishing core collections (subsets of accessions designed to capture the greatest amount of genetic diversity) and detecting admixed populations (e.g., Myles et al., 2011).

Wild sunflowers have long been used in sunflower improvement (Korell et al., 1992; Seiler, 1992; Baute et al., 2015). However, use of wild species has been hampered by a lack of high-resolution characterization of genotypic diversity maintained in public seed banks (Smith, 2015). The genus contains circa 12 annual and 37 perennial species distributed across North America in ecologically diverse habitats that range from open plains to salt marshes (Heiser et al., 1969; Rogers et al., 1982; Schilling, 2006). Reconstructing phylogenetic relationships among these species has been a formidable

[1] Manuscript received 9 August 2016; revision accepted 10 November 2016.
[2] Department of Botany, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada; and
[3] Department of Biology, Indiana University, Bloomington, Indiana 47405 USA
[4] Author for correspondence (e-mail: loren.rieseberg@botany.ubc.ca)
*Co-first authors: contributed equally.
doi:10.3732/ajb.1600295

challenge due to the group's recent origin (Schilling, 1997), high incidence of interspecific hybridization (Kane et al., 2009), and occurrence of multiple rounds of whole-genome duplication (Barker et al., 2008). These same factors have also made it difficult to resolve species boundaries among some *Helianthus* taxa. Although recent work suggests that next-generation sequence data can be used to resolve some of these relationships (Bock et al., 2014; Stephens et al., 2015), the phylogenetic position of the most newly identified *Helianthus* species, *H. winteri* J.C.Stebbins, has not been formally resolved (Stebbins et al., 2013).

Detailed genetic characterization is needed at the intraspecific level as well. Pinpointing genetic population structure serves the interests of breeders who want to capture the most diversity with the fewest samples. Furthermore, genetic structure may also correspond to important phenotypic traits (e.g., oil content) or environmental adaptations (e.g., drought tolerance or avoidance) (Seiler, 1992). The wild progenitor of the cultivated sunflower, also *H. annuus*, occurs across much of North America (Heiser et al., 1969) and is known to have a large effective population size (Strasburg et al., 2011). Previous work has identified subpopulation structure within *H. annuus*, corresponding to the divergence of populations in California (Dorado et al., 1992) and Texas (Rieseberg et al., 1990). Other than these populations, little genetic structure has been reported previously within *H. annuus* (Mandel et al., 2011). Coincident with the geographic subpopulations, hybridization and introgression with congeners has been shown in California, with *H. bolanderi-exilis* (Owens et al., 2016), and in Texas, with *H. debilis* (Rieseberg et al., 1990). This coincidence raises the intriguing possibility that introgression is driving the observed population structure, similar to patterns found in some domesticate–wild introgression (Magnussen and Hauser, 2007; Goedbloed et al., 2013). Answering this question is important. If population structure is entirely driven by introgression, geographic subpopulations offer only new variation from congeners. Alternatively, if population structure is independent of introgression, then geographic subpopulations can contribute their own unique diversity not found in congeners.

Here, we used genotyping by sequencing (GBS) to survey genome-wide genetic variation in 292 accessions of wild *Helianthus* from 22 taxa. We used these data to reconstruct phylogenomic relationships among annual and perennial sunflowers. We then investigated patterns of genetic diversity within all sampled *Helianthus* species and population structure within a geographically diverse panel of wild *H. annuus* accessions. We used the ABBA-BABA or Patterson's *D* statistic approach (Kulathinal et al., 2009; Green et al., 2010; Durand et al., 2011) to quantify gene flow between *H. annuus* and its sympatric annual relatives across its range. Finally, we investigated whether the detected patterns of population structure are a result of introgression between these species using a novel method that asks whether the loci responsible for principal component (PC) separation are biased toward a potential donor species.

MATERIALS AND METHODS

The USDA gene bank contains more than 2000 accessions of wild annual *Helianthus* (Kane et al., 2013). We selected approximately 15% of these accessions for genetic analyses. This selection focused on covering the indigenous geographic range of *H. annuus* (Rogers et al., 1982). In addition, we included representatives from each species of the annual clade of sunflowers, as well several cross-compatible perennial species (Appendices S1 and S2, see Supplemental Data with the online version of this article). Lastly, to survey species that were not available from the USDA at the time, for example, *H. winteri*, we included some Rieseberg laboratory collections (Appendix S2).

DNA was isolated from leaf tissue using a CTAB protocol (Doyle and Doyle, 1987). We genotyped these samples using a modified GBS protocol (Elshire et al., 2011). Briefly, we digested DNA for each sample with PstI, ligated fragments to Illumina sequencing adapters and custom barcodes and amplified pooled samples with 18 cycles of polymerase chain reaction (PCR). In contrast to the original protocol, a gel electrophoresis/isolation step was incorporated to exclude dimers amplified during PCR. Ninety-six samples were sequenced per lane on an Illumina HiSeq 2000 (Illumina, San Diego, California) using paired-end sequencing. The reads were demultiplexed using an in-house Perl script that also removed adapter read-through. Reads shorter than 50 bp following the trimming step were removed. Reads were trimmed for base quality and Illumina adapter using Trimmomatic (v0.32) and the options "ILLUMINACLIP: TruSeq3-PE.fa:2:30:10:8:T SLIDINGWINDOW:4:15 MINLEN:36" (Bolger et al., 2014), and the remaining reads were aligned to a genome assembly of *H. annuus* (v1.1.bronze; http://www.sunflowergenome.org) using the program NextGenMap (v0.4.12, Sedlazeck et al., 2013). Variants were called using the program FreeBayes (v1.0.2, Garrison and Marth, 2012). Although some species in our data set are polyploid, we treated all samples as diploid for the FreeBayes step. This approach allowed us to use joint variant calling to maximize variant discovery and facilitated downstream analyses since many tools do not yet accommodate polyploid single nucleotide polymorphism (SNP) calls. Analyses that may have been affected by ploidy variation are noted. All custom scripts and a filtered fasta used for phylogenetics can be found on github (https://github.com/owensgl/Helianthus_diversity). The raw sequence data are stored in the Sequence Read Archive (SRA, #SRP062491, https://www.ncbi.nlm.nih.gov/sra).

To simplify analyses, we removed all indels and multinucleotide variants, leaving only SNPs and invariant sites. Individual genotypes were filtered to remove calls with <5 reads. For diversity analytics, all sites were used. We calculated Nei's genetic diversity, or average expected heterozygosity (Nei, 1973). For this, we did not include samples whose species identity did not match with results from phylogenetic analyses (see below). Species level variance was estimated by jackknife resampling each species.

For all phylogenetic analyses, we selected SNPs that passed two filtering criteria: minor allele frequency > 0.01 and missing data < 20%. A phylogenetic network was generated with the program Splitstree4 (Huson, 1998) using default settings. A maximum likelihood tree was created using IQtree software (Nguyen et al., 2015), and confidence was assessed using 1000 ultrafast bootstrap approximation replicates and 1000 bootstrap SH-like approximate likelihood ratio test replicates. The best substitution model was determined using the "–m TEST" option in IQtree, which evaluates different substitution models and selects the best model using the Bayesian information criterion. To account for the ascertainment bias of only using variable sites, we used the "+ASC" option. To assess population structure, we used a principal component analysis (PCA) using the R package SNPrelate (Zheng et al., 2012). Loci were trimmed to linkage disequilibrium (LD <0.2) using the command

"snpgdsLDpruning". We also set a threshold for minor allele frequency >0.05.

To test for localized gene flow with *H. annuus* and other annual species, we ran ABBA-BABA tests (Kulathinal et al., 2009; Green et al., 2010; Durand et al., 2011). The ABBA-BABA test, also known as Patterson's *D* statistic, uses a simple phylogenetic framework to determine whether a population's genetic composition has been influenced by interspecific gene flow. It looks for loci where the gene tree does not agree with the species tree in a four member phylogeny (labeled P1 through P4) with the pattern (((P1, P2),P3),P4). In this species phylogeny, we expect derived alleles to be shared by P1 and P2, or P1, P2 and P3. Incomplete lineage sorting will produce equal numbers of loci where P1 and P3 share a derived allele (i.e., BABA) and loci where P2 and P3 share a derived allele (i.e., ABBA). In contrast, gene flow with P3 will produce an excess of BABAs or ABBAs if P1 or P2 respectively, is involved. In our analysis, we placed allopatric *H. annuus* in position P1, sympatric *H. annuus* in P2, and the potential hybridizing species in P3. We assigned each sample to one of three groups: California (sympatric with *H. bolanderi-exilis*), Texas (sympatric with *H. argophyllus*, *H. debilis*, and *H. praecox*) and central USA (not sympatric with any tested species). We chose these regions because previous work has found evidence of gene flow from *H. bolanderi-exilis* into Californian *H. annuus* and from *H. debilis* into Texan *H. annuus* (Rieseberg et al., 1990; Owens et al., 2016). Furthermore, the first two PCs separate California and Texas from the rest of the samples. For the ABBA-BABA analysis, we filtered for sites that had representation in each group, were biallelic, and were fixed for a single allele in all ancestral taxa. We treated a pooled group of all perennial species (*H. decapetalus*, *H. giganteus*, *H. grosseserratus*, *H. hirsutus*, *H. maximilliani*, *H. nuttallii*, *H. strumosus*, and *H. tuberosus*) as our ancestral clade to ensure that ancestral alleles were correctly identified. For *H. bolanderi-exilis*, we treated California as sympatric and Texas and central as allopatric. For *H. argophyllus*, *H. debilis*, and *H. praecox*, we treated Texas as sympatric and California and central as allopatric. Because mislabelled samples can have a large effect, we only included individuals for which the sample identity matched the genetic assignment in the phylogenetic analysis. Standard errors were calculated using a block jackknife bootstrap with 10 MB as the block size. We also calculated f_d (Martin et al., 2013) to estimate the proportion of the genome shared through introgression and estimated standard errors in the same way as for the *D* statistic. Note that we did not test for introgression with *H. petiolaris*. *Helianthus petiolaris* is sympatric with *H. annuus* over much of its central range, and other work has suggested ongoing gene flow; thus, it is impossible to select meaningful allopatric populations of *H. annuus*.

Lastly, we investigated whether episodes of introgression are driving the detected patterns of *H. annuus* population structure using a novel method. If gene flow is driving population structure, then we would expect that for the SNPs with the highest loading value (i.e., the SNPs most responsible for the PC separation), the potential gene flow donor species would have a closer allele frequency to the *H. annuus* samples that correspond to the extreme PC values. In other words, at the loci responsible for separating California *H. annuus* from the rest of samples, *H. bolanderi-exilis* should be more similar to California *H. annuus* than the rest of *H. annuus* if introgression is causing the PCA separation. Similarly, we expected the same pattern for *H. argophyllus* and PC2. We extracted the SNP identity of the PCA loadings for the first two PCs. For the SNPs with the top 100 absolute loading values, we calculated the allele frequency for the candidate gene flow donor (i.e., *H. bolanderi-exilis* and *H. argophyllus*). We also calculated the allele frequency for *H. annuus* samples with a positive score in the PC and for *H. annuus* samples with a negative score in the PC. We asked whether the allele frequency for the donor species was closer to the allele frequency of the positive or negative group and used a binomial test for deviation from equality. If introgression is not involved, we expect the donor species to be more similar to positive and negative PC groups in equal proportions. For PC2, we also repeated the analysis with *H. debilis* and *H. praecox* as possible donors.

RESULTS

We sequenced 292 samples representing 22 taxa of *Helianthus*. Following de-multiplexing, four of these samples were removed due to insufficient sequencing depth (Appendix S1). A total of 3,179,885 sites were genotyped in at least one individual, including invariant sites. After filtering for >80% call rate and >1% minor allele frequency, 4645 SNPs remained for analyses.

In our phylogenetic analyses, while the majority of samples clustered with other samples of the same taxon as expected, there were 14 exceptions (Fig. 1; online Appendices S2, S3). These outlier samples typically corresponded to accessions with collection notes suggestive of questionable origins. For example, the collection notes for a *H. petiolaris* sample found in the *H. annuus* cluster states "*Helianthus annuus* plants mixed with *H. petiolaris* on both sides of road". An outlier sample annotated as a *H. maximiliani* that also clusters with *H. annuus* was selected from a wild collection for being "bigger" than the rest of the *H. maximiliani*. Because outliers likely represent misclassification, we excluded them from further analyses. Most named taxa are supported as unique lineages by these trees, with the exception of *H. petiolaris*/*H. neglectus*, *H. bolanderi*/*H. exilis*, *H. winteri*/*H. annuus*, *H. hirsutus*/*H. divaricatus*/*H. strumosus*, and *H. tuberosus*.

We analyzed genetic diversity for each species and population structure for *H. annuus*. Genetic diversity was highest in *H. neglectus*, *H. tuberosus*, *H. annuus*, and *H. petiolaris*, and lowest in *H. paradoxus*, *H. winteri*, and *H. nuttallii* (Fig. 2; online Appendix S4). Analyses of population structure within *H. annuus* were based on 100 samples from across the geographic range of the species and 774 unlinked SNPs. The first four PCs indicated geographic population structure, although total percentage variance explained was low (5.5–2.2% for PC1 to PC4). The first PC separated California samples from the rest of the samples. The second PC separated samples generally by latitude and highlighted southern Texas samples as outliers. The third PC also separated populations by latitude generally and the fourth PC separated central New Mexico and Arizona populations from the rest of the range (Fig. 3; online Appendices S5, S6).

Tests for localized introgression between *H. annuus* and sympatric populations of other annual sunflowers detected a significant signal for *H. bolanderi-exilis* and *H. argophyllus*; 9.7 ± 1.8% and 7.1 ± 3.2% of the genome, respectively, was shown to be involved in introgression using the f_d statistic. All other comparisons did not reach our threshold for significance (Fig. 4, online Appendix S7).

Because the first two PCs highlighted California and Texas, both regions where interspecific gene flow has been reported to occur, we tested whether introgression was responsible for the PCs. We

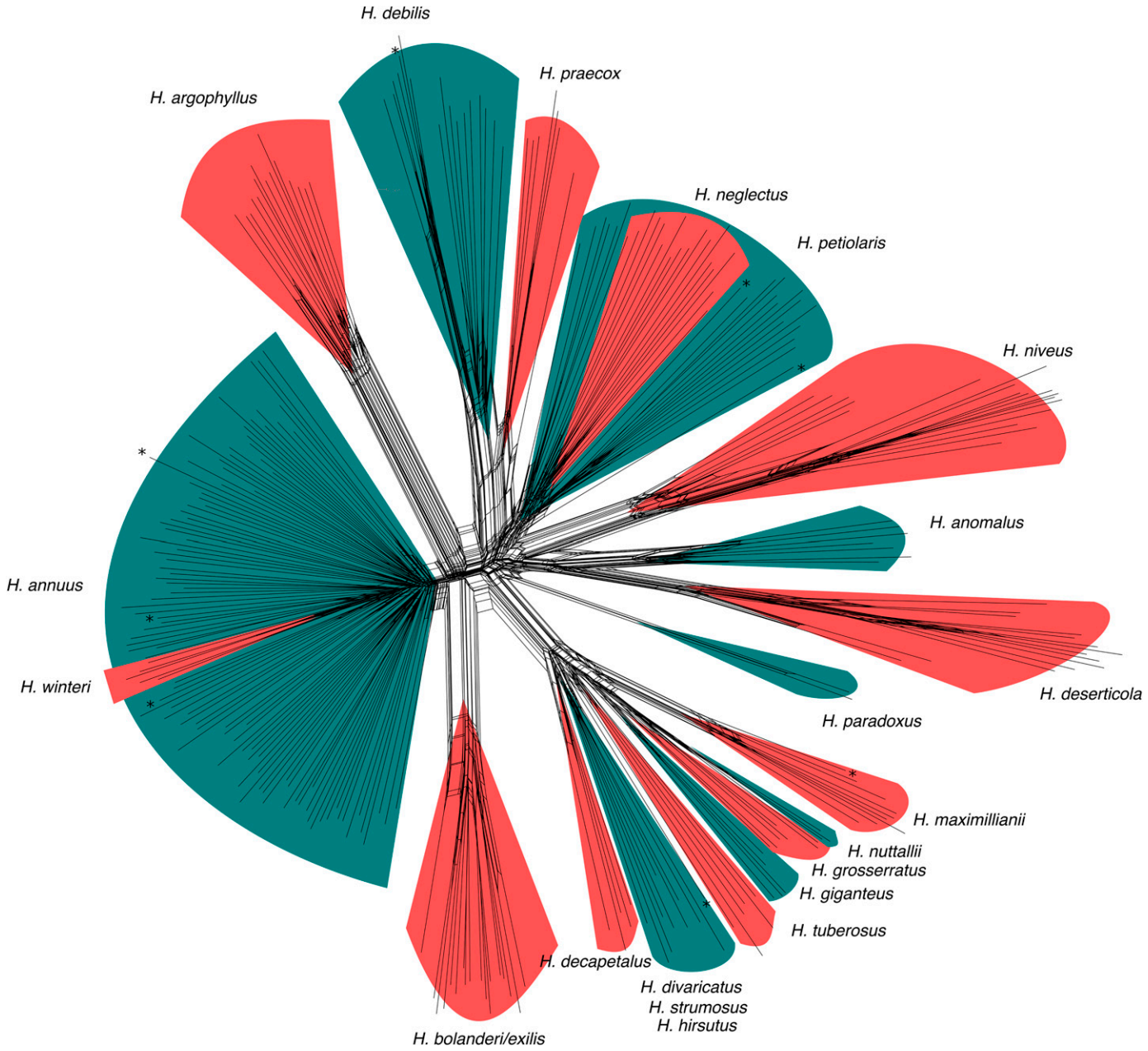

FIGURE 1 Phylogenetic network of *Helianthus* germplasm made using Splitstree4. Samples whose genetic clustering does not match germplasm species ID are highlighted with an asterisk (*).

found that for PC1, for the top 100 loading sites *H. bolanderi-exilis* more often had a closer allele frequency to the negative PC score group (i.e., the rest of the range) than the positive PC score group (i.e., California) (binomial test, measured = 33%, expected = 50%, $p < e^{-5}$). For the second PC at the top 100 loading sites, when examining allele frequencies for *H. argophyllus* (measured = 50%, expected = 50%, $p = 1$), *H. debilis* (measured = 43%, expected = 50%, $p = 0.21$), and *H. praecox* (measured = 46%, expected = 50%, $p = 0.47$), we failed to find a deviation from the neutral expectation. Thus, they were equally likely to be similar to the positive PC score group (i.e., the northern half of the range) and the negative PC score group (i.e., Texas). All together, we failed to find any signal that introgression is driving population structure in *H. annuus*.

DISCUSSION

By surveying numerous samples across a wide variety of species in *Helianthus* with dense genetic markers, we have made two key advances: (1) we provide unambiguous species identification for a majority of accessions genotyped from a public seed bank and highlight the ambiguous genetic relationships of the rest, and (2) we show regional genetic variation in wild *H. annuus* including localized introgression from close relatives. Both of these advances will be useful to breeding efforts that seek to use wild diversity in the most efficient manner.

Phylogenetics and diversity of Helianthus—Accurate species identification in germplasm collections is critical for breeders and other

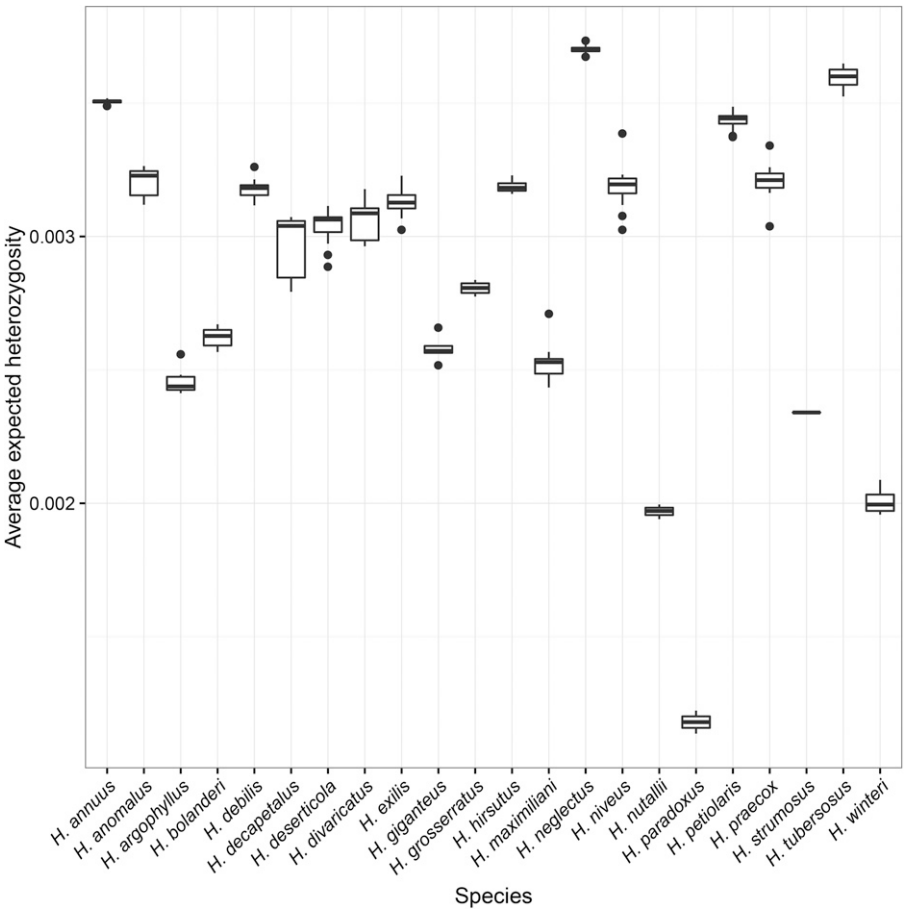

FIGURE 2 Nei's genetic diversity for each species. Error bars indicate variance measured from jackknife resampling.

end-users of germplasm. Analysis of 292 samples genotyped for thousands of markers with the GBS approach revealed several instances where genotypic and collection information support each other in highlighting specific collections that appear to have been misidentified. For example, a *H. petiolaris* sample collected near a population of *H. annuus* was found to cluster genetically with *H. annuus* (Fig. 1). Another example involves several accessions that were originally identified as *H. bolanderi* but appear genetically to be *H. annuus*. These samples were found to resemble *H. annuus* phenotypically as well (data not shown). This information has already been incorporated into the USDA germplasm information database, and the collections have been reclassified (Appendix S2). Although the sampling described here is far from comprehensive, even for the subset of taxa investigated, it does demonstrate the power of using genomic tools for addressing practical issues concerning germplasm curation.

Our phylogenetic analyses based on GBS data largely agree with the recent *Helianthus* phylogeny produced by Stephens et al. (2015) with a few extra details. With greater individual sampling, we can more confidently say that populations identified as *H. neglectus* largely group together with the *H. petiolaris* subsp. *fallax* clade but that the grouping does not exclude all *H. petiolaris* samples. Similarly, *H. winteri*, identified as being closely related to *H. annuus* previously, groups within the larger *H. annuus* clade (Fig. 1) (Stebbins et al., 2013). This evidence suggested that *H. winteri* is a young species and that it originated after *H. annuus* spread across its range or has had extensive hybridization with *H. annuus*. As found in previous work, we did not detect consistent genetic differences between *H. bolanderi* and *H. exilis* (Owens et al., 2016).

With a few notable exceptions, the named taxa are well supported as independent lineages. The hybrid species, *H. paradoxus*, is placed at the base of the annual clade, possibly due to its relatively balanced hybrid ancestry (Rieseberg, 2003). The other two hybrid species, *H. deserticola* and *H. anomalus*, are placed within the *H. petiolaris* clade, supporting previous morphological analyses that suggested they are more similar to *H. petiolaris* (Rosenthal et al., 2002). Despite considerable genetic distance from the reference sequence and variation in ploidy, the perennial samples are well resolved, and most taxa are supported. There is no differentiation between the autotetraploid *H. hirsutus* and its diploid progenitor, *H. divaricatus*. Similarly, with one exception, diploid and autotetraploid *H. decapetalus* form a monophyletic clade. As expected, the placement of *H. tuberosus*, an autoallohexaploid, is unresolved, but in a clade with its diploid and tetraploid parents, *H. grosseserratus* and *H. hirsutus* (Bock et al., 2014), among other species. Thus, while we did not specifically attempt to address genotyping problems that might arise from polyploidy, the placement we recovered for this polyploid is congruent with its known ancestry (Bock et al., 2014).

We found highly variable amounts of diversity among species. High diversity in *H. neglectus* is surprising considering its restricted species range but is consistent with previous studies (Raduski et al., 2010). This diversity may reflect ongoing gene flow with a related species. *Helianthus tuberosus* is a polyploid hybrid, which was not accounted for in SNP calling, so high diversity is most likely at least partially due to the extra genomic content. Both *H. annuus* and *H. petiolaris*, the two annual *Helianthus* species with the largest ranges, have high diversity as expected. The lowest diversity was seen in *H. paradoxus*, *H. winteri*, and *H. nuttallii*, three species with relatively restricted ranges.

***Population structure of* H. annuus**—The progenitor of the cultivated sunflower, *H. annuus*, contains a striking amount of diversity (Fig. 2). Although it is thought to be largely homogeneous across its range (Mandel et al., 2011), we found some structure within *H. annuus* consistent with isolation by distance, although total percentage variance explained for the first four PCs was 13.1% (Fig. 3). Preliminary analysis with fastSTRUCTURE, not presented here, found variable subpopulation assignment between runs and between data subsets, although this analysis consistently suggested a K value > 1. This result is likely because the isolation by distance pattern without sharp divides between subpopulations is not a good fit for the fastSTRUCTURE model. Although we discuss subpopulations below, the low percentage variance suggests a

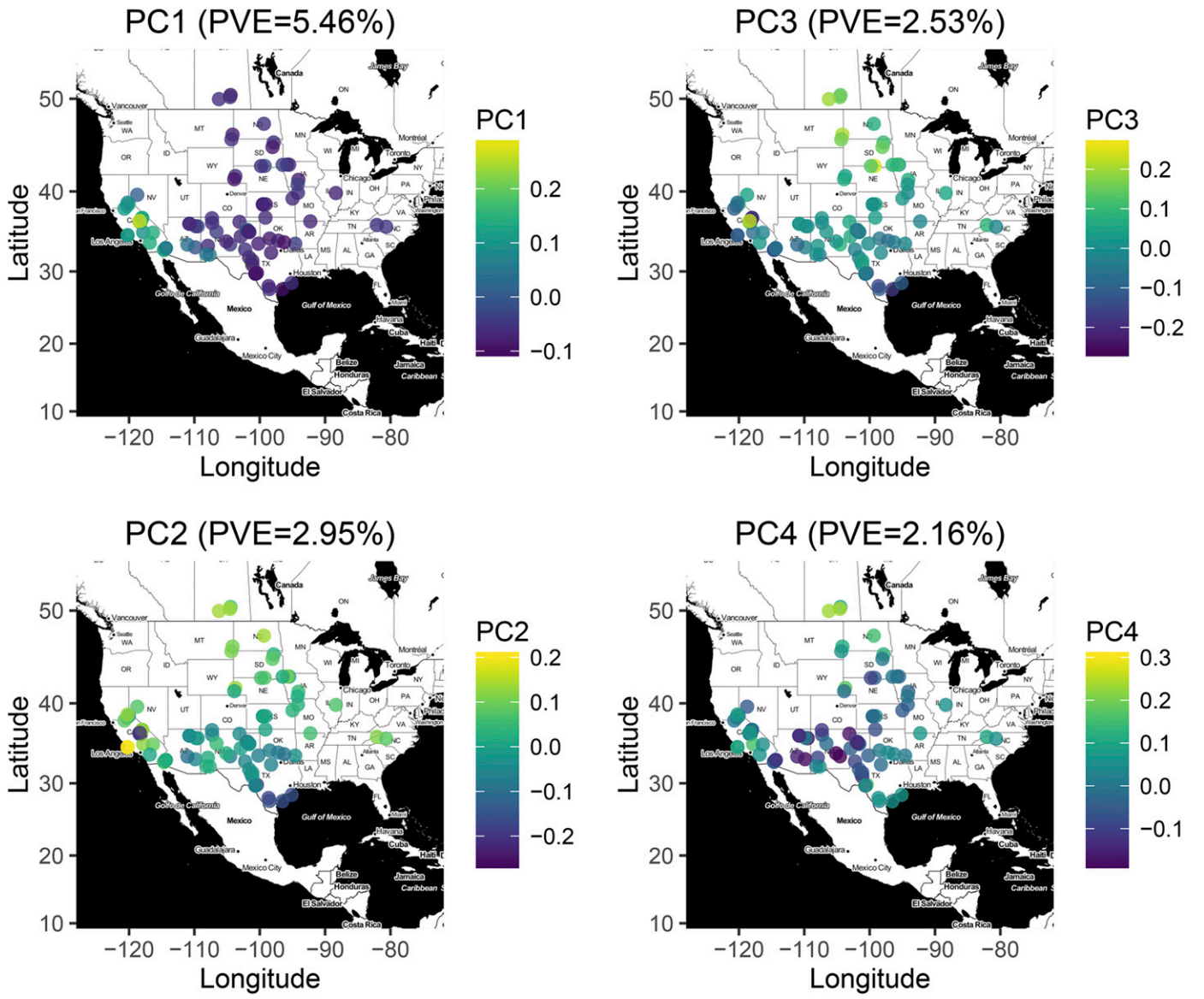

FIGURE 3 Principal component score for *Helianthus annuus* samples, plotted by geographic location for the first four PCs. Color represents the PC score for each sample.

well-mixed metapopulation structure with gradual isolation by distance over long geographic distances, with the possible exception of the California subpopulation.

The California subpopulation is consistent with genetic drift after population expansion. Charles Heiser suggested that *H. annuus* invaded California after being brought there by Native Americans before Europeans arrived in North America (Heiser, 1949). Interestingly, although we find evidence of gene flow with *H. bolanderi-exilis*, and previous work has suggested that the gene flow was primarily into *H. annuus* (Owens et al., 2016), at the SNPs most responsible for separating PC1, *H. bolanderi-exilis* is more similar to the non-Californian *H. annuus*. One scenario that explains this surprising pattern is that Californian *H. annuus* underwent a bottleneck during invasion and experienced additional genetic drift and shifts in allele frequency. These shifts would have pushed the allele frequencies further from the allele frequency of the ancestor of both *H. annuus* and *H. bolanderi-exilis*. Thus, at the high loading sites, *H. bolanderi-exilis* is more similar to the non-Californian *H. annuus* population because it has undergone less drift and is closer to the ancestral state.

The second PC separated Texas samples from the rest, but also more generally divided samples on a latitudinal gradient. This finding is somewhat consistent with the range of the subspecies *H. annuus* subsp. *texanus*, which is thought to contain introgression from *H. debilis* (Heiser, 1951; Rieseberg et al., 1990; Scascitelli et al., 2010). Despite this, we found evidence of gene flow with *H. argophyllus* in Texas but not with *H. debilis* using the ABBA-BABA test. By examining the loci responsible for PC separation, we failed to find a signal that introgression from any sympatric species is driving the population structure pattern found in Texas. This result should caution future research from ascribing population structure to introgression, even when introgression and population structure colocalize geographically.

We examined the highest PC loading loci to determine whether introgression was causing signals in the PCA. In this case, we found

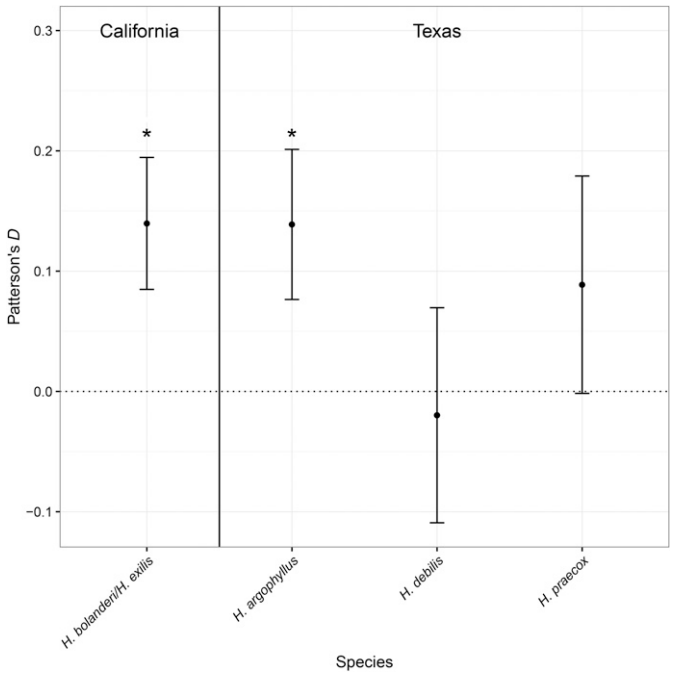

FIGURE 4 Interspecific gene flow between sympatric *Helianthus annuus* and local annual *Helianthus* determined by ABBA BABA tests. More positive D-values indicate stronger evidence of introgression. The asterisk indicates contrasts with a significant difference from zero ($p < 0.05$); error bars indicate standard error. Full results are reported in Appendix S7.

either no signal or a signal in the opposite direction than expected. We have attributed this pattern to a population bottleneck, but more broadly it suggests that this sort of test may be prone to errors due to underlying demographic processes. Thus, although we feel our results support population structure independent of introgression, the potential for false positives suggests the need for caution when using this or similar approaches. More work needs to be done on the question of introgression-derived population structure to determine robust statistical tests.

***Gene flow with* H. annuus**—*Helianthus annuus* is well known to hybridize with the wide-ranging *H. petiolaris*, but our work suggests it is also hybridizing with its two closest relatives *H. argophyllus* and *H. bolanderi-exilis* (Fig. 4) (Rieseberg et al., 1998; Strasburg and Rieseberg, 2008). We did not test for introgression with *H. petiolaris* because of a lack of plausible nonintrogressed allopatric populations. Previous work has found evidence the Texas subspecies, *H. annuus* subsp. *texanus*, is a product of introgression with *H. debilis* (Rieseberg et al., 1990; Whitney et al., 2006, 2010; Scascitelli et al., 2010). Surprisingly, we failed to detect a signal of this proposed event. This failure may be due to widespread introgression with *H. petiolaris* masking the signal of *H. debilis* introgression. To understand why, we can picture a scenario where P1 and P3 had gene flow, but P2 and P3 also had gene flow. The first would produce extra BABA loci, and the second extra ABBA loci, but because the ABBA-BABA test looks for enrichment of one pattern, it would not find a signal if they balanced out. In our work, *H. petiolaris* is closely related to *H. debilis*; therefore, introgression from *H. petiolaris* into central USA *H. annuus* may mask localized introgression from *H. debilis* in a similar manner.

Understanding patterns of introgression in the wild could be of use to crop improvement programs. For example, introgression associated with the colonization of new habitats could be adaptive and involve genetic transfer of alleles of interest for crop improvement. Conversely, genomic regions that resist introgression could harbor locally adapted alleles and thus represent potential targets for further investigation. Admixed wild genotypes could also be used as a bridge to help breeders access a larger amount of genetic diversity with less effort. For example, it may be possible to introgress *H. bolanderi-exilis* alleles into cultivated *H. annuus* via wild *H. annuus* from California without suffering the cost of the reduced hybrid fertility. In another example, modern cultivars of *H. annuus* appear to contain alleles from the Texas subpopulation of *H. annuus*, *H. annuus* subsp. *texanus* (Baute et al., 2015). Although our current analysis failed to find evidence for the hybrid ancestry of *H. annuus* subsp. *texanus*, it may be acting as a bridge for *H. debilis* alleles. It is not uncommon for breeders to employ such a bridge process to move alleles into cultivars of interest (Jansky and Hamernik, 2009). Using genomic data to identify introgression in the wild could allow breeders to access more distant wild relatives more rapidly and efficiently.

ACKNOWLEDGEMENTS

This work was undertaken as part of the initiative "Adapting Agriculture to Climate Change: Collecting, Protecting and Preparing Crop Wild Relatives" which is supported by the Government of Norway. The project is managed by the Global Crop Diversity Trust with the Millennium Seed Bank of the Royal Botanic Gardens, Kew and implemented in partnership with national and international gene banks and plant breeding institutes around the world. For further information, see the project website: http://www.cwrdiversity.org/. This research was also supported by Genome Canada, Genome BC, and NSERC Fellowships to G. Baute, G. Owens and D. Bock.

CONFLICTS OF INTEREST

We declare no conflicts of interest.

LITERATURE CITED

Barker, M. S., N. C. Kane, M. Matvienko, A. Kozik, R. 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.

Baute, G. J., N. C. Kane, C. J. Grassa, Z. Lai, and L. H. Rieseberg. 2015. Genome scans reveal candidate domestication and improvement genes in cultivated sunflower, as well as post-domestication introgression with wild relatives. *New Phytologist* 206: 830–838.

Bock, D. G., N. C. Kane, D. P. Ebert, and L. H. Rieseberg. 2014. Genome skimming reveals the origin of the Jerusalem artichoke tuber crop species: Neither from Jerusalem nor an artichoke. *New Phytologist* 201: 1021–1030.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120.

Dorado, O., L. H. Rieseberg, and D. M. Arias. 1992. Chloroplast DNA introgression in southern California sunflowers. *Evolution* 46: 566–572.

Doyle, J. J., and J. L. Doyle. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 19: 11–15.

Durand, E. Y., N. Patterson, D. Reich, and M. Slatkin. 2011. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution* 28: 2239–2252.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6: e19379.

Garrison, E., and G. Marth. 2012. Haplotype-based variant detection from short-read sequencing [version 2]. arXiv:1207.3907v2 [q-bio.GN].

Goedbloed, D. J., P. van Hooft, H. J. Megens, K. Langenbeck, W. Lutz, R. P. M. A. Crooijmans, S. E. van Wieren, et al. 2013. Reintroductions and genetic introgression from domestic pigs have shaped the genetic population structure of Northwest European wild boar. *BMC Genetics* 14: 43.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, et al. 2010. A draft sequence of the Neandertal genome. *Science* 328: 710–722.

Hajjar, R., and T. Hodgkin. 2007. The use of wild relatives in crop improvement: A survey of developments over the last 20 years. *Euphytica* 156: 1–13.

Heiser, C. B. 1949. Study in the evolution of the sunflower species *Helianthus annuus* and *H. bolanderi*. *University of California Publications in Botany* 23: 157–196.

Heiser, C. B. 1951. Hybridization in the annual sunflowers: *Helianthus annuus* X *H. debilis* var. *cucumerifolius. Evolution* 5: 42–51.

Heiser, C. B., D. M. Smith, S. B. Clevenger, and W. C. Martin. 1969. The North American sunflowers (*Helianthus*). *Memoirs of the Torrey Botanical Club* 22: 1–218.

Huson, D. H. 1998. SplitsTree: Analyzing and visualizing evolutionary data. *Bioinformatics* 14: 68–73.

Jansky, S., and A. Hamernik. 2009. The introgression of 2× 1EBN *Solanum* species into the cultivated potato using *Solanum verrucosum* as a bridge. *Genetic Resources and Crop Evolution* 56: 1107–1115.

Kane, N. C., J. M. Burke, L. Marek, G. Seiler, F. Vear, G. Baute, S. J. Knapp, et al. 2013. Sunflower genetic, genomic and ecological resources. *Molecular Ecology Resources* 13: 10–20.

Kane, N. C., M. G. King, M. S. Barker, A. Raduski, S. Karrenberg, Y. Yatabe, S. J. Knapp, and L. H. Rieseberg. 2009. Comparative genomic and population genetic analyses indicated highly porous genomes and high levels of gene flow between divergent *Helianthus* species. *Evolution* 63: 2061–2075.

Korell, M., G. Mösges, and W. Friedt. 1992. Construction of a sunflower pedigree map. *Helia* 15: 7–16.

Kulathinal, R. J., L. S. Stevison, and M. A. F. Noor. 2009. The genomics of speciation in *Drosophila*: Diversity, divergence, and introgression estimated using low-coverage genome sequencing. *PLOS Genetics* 5: e1000550.

Magnussen, L. S., and T. P. Hauser. 2007. Hybrids between cultivated and wild carrots in natural populations in Denmark. *Heredity* 99: 185–192.

Mandel, J. R., J. M. Dechaine, L. F. Marek, and J. M. Burke. 2011. Genetic diversity and population structure in cultivated sunflower and a comparison to its wild progenitor, *Helianthus annuus* L. *Theoretical and Applied Genetics* 123: 693–704.

Martin, S. H., K. K. Dasmahapatra, N. J. Nadeau, C. Salazar, J. R. Walters, F. Simpson, M. Blaxter, et al. 2013. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Research* 23: 1817–1828.

McCouch, S., G. J. Baute, J. Bradeen, P. Bramel, P. K. Bretting, E. Buckler, J. M. Burke, et al. 2013. Agriculture: Feeding the future. *Nature* 499: 23–24.

Myles, S., A. R. Boyko, C. L. Owens, P. J. Brown, F. Grassi, M. K. Aradhya, B. Prins, et al. 2011. Genetic structure and domestication history of the grape. *Proceedings of the National Academy of Sciences, USA* 108: 3530–3535.

Nei, M. 1973. Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences, USA* 70: 3321–3323.

Nguyen, L.-T., H. A. Schmidt, A. von Haeseler, and B. Q. Minh. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* 32: 268–274.

Owens, G. L., G. J. Baute, and L. H. Rieseberg. 2016. Revisiting a classic case of introgression: Hybridization and gene flow in Californian sunflowers. *Molecular Ecology* 25: 2630–2643.

Raduski, A. R., L. H. Rieseberg, and J. L. Strasburg. 2010. Effective population size, gene flow, and species status in a narrow endemic sunflower, *Helianthus neglectus*, compared to its widespread sister species, *H. petiolaris*. *International Journal of Molecular Sciences* 11: 492–506.

Rieseberg, L. H. 2003. Major ecological transitions in wild sunflowers facilitated by hybridization. *Science* 301: 1211–1216.

Rieseberg, L. H., S. J. E. Baird, and A. M. Desrochers. 1998. Patterns of mating in wild sunflower hybrid zones. *Evolution* 52: 713–726.

Rieseberg, L. H., S. Beckstrom-Sternberg, and K. Doan. 1990. *Helianthus annuus* ssp. *texanus* has chloroplast DNA and nuclear ribosomal RNA genes of *Helianthus debilis* ssp. *cucumerifolius. Proceedings of the National Academy of Sciences, USA* 87: 593–597.

Rogers, C. E., T. E. Thompson, and G. J. Seiler. 1982. Sunflower species of the United States. National Sunflower Association, Bismarck, North Dakota, USA.

Rosenthal, D. M., A. E. Schwarzbach, L. A. Donovan, O. Raymond, and L. H. Rieseberg. 2002. Phenotypic differentiation between three ancient hybrid taxa and their parental species. *International Journal of Plant Sciences* 163: 387–398.

Scascitelli, M., K. D. Whitney, R. A. Randell, M. King, C. A. Buerkle, and L. H. Rieseberg. 2010. Genome scan of hybridizing sunflowers from Texas (*Helianthus annuus* and *H. debilis*) reveals asymmetric patterns of introgression and small islands of genomic differentiation. *Molecular Ecology* 19: 521–541.

Schilling, E. E. 1997. Phylogenetic analysis of *Helianthus* (Asteraceae) based on chloroplast DNA restriction site data. *Theoretical and Applied Genetics* 94: 925–933.

Schilling, E. E. 2006. *Helianthus*. *In* Flora of North America Editorial Committee [eds.], Flora of North America north of Mexico, vol. 21, 141–169. Oxford University Press, New York, New York, USA.

Sedlazeck, F. J., P. Rescheneder, and A. von Haeseler. 2013. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* 29: 2790–2791.

Seiler, G. J. 1992. Utilization of wild sunflower species for the improvement of cultivated sunflower. *Field Crops Research* 30: 195–230.

Smith, C. 2015. What the wild things do: The use of crop wild relatives in public international breeding programs and implications for conservation. M.Sc. thesis, University of Waterloo, Waterloo, Ontario, Canada.

Stebbins, J., C. J. Winchell, and J. V. H. Constable. 2013. *Helianthus winteri* (Asteraceae), a new perennial species from the southern Sierra Nevada foothills, California. *Aliso* 31: 19–23.

Stephens, J. D., W. L. Rogers, C. M. Mason, L. A. Donovan, and R. L. Malmberg. 2015. Species tree estimation of diploid *Helianthus* (Asteraceae) using target enrichment. *American Journal of Botany* 102: 910–920.

Strasburg, J. L., N. C. Kane, A. R. Raduski, A. Bonin, R. Michelmore, and L. H. Rieseberg. 2011. Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. *Molecular Biology and Evolution* 28: 1569–1580.

Strasburg, J. L., and L. H. Rieseberg. 2008. Molecular demographic history of the annual sunflowers *Helianthus annuus* and *H. petiolaris*—Large effective population sizes and rates of long-term gene flow. *Evolution* 62: 1936–1950.

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.

Zheng, X., D. Levine, J. Shen, S. M. Gogarten, C. Laurie, and B. S. Weir. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28: 3326–3328.