MOLECULAR ECOLOGY

Molecular Ecology (2016) **25**, 2630–2643 doi: 10.1111/mec.13569

GENOMICS OF HYBRIDIZATION

Revisiting a classic case of introgression: hybridization and gene flow in Californian sunflowers

GREGORY L. OWENS,* GREGORY J. BAUTE* and LOREN H. RIESEBERG*†

Department of Botany and Biodiversity Research Centre, University of British Columbia, University Blvd, Vancouver, BC V6T 1Z4, Canada, †Department of Biology, Indiana University, Bloomington, IN 47405, USA

Abstract

During invasion, colonizing species can hybridize with native species, potentially swamping out native genomes. However, theory predicts that introgression will often be biased into the invading species. Thus, empirical estimates of gene flow between native and invasive species are important to quantify the actual threat of hybridization with invasive species. One classic example of introgression occurs in California, where *Helianthus bolanderi* was thought to be a hybrid between the serpentine endemic *Helianthus exilis* and the congeneric invader *Helianthus annuus*. We used genotyping by sequencing to look for signals of introgression and population structure. We find that *H. bolanderi* and *H. exilis* form one genetic clade, with weak population structure that is associated with geographic location rather than soil composition and likely represent a single species, not two. Additionally, while our results confirmed early molecular analysis and failed to support the hybrid origin of *H. bolanderi*, we did find evidence for introgression mainly into the invader *H. annuus*, as predicted by theory.

Keywords: angiosperms, gene flow, hybridization, invasive species, phylogeography, population genetics—empirical

Received 22 September 2015; revision received 26 January 2016; accepted 28 January 2016

Introduction

In Verne Grant's seminal work 'Plant Speciation', he lists four examples of introgression, one of which involves the sunflower *Helianthus bolanderi* (Grant 1981). Both morphology and habitat suggested that this largely ruderal species was a product of introgression between the smaller native serpentine endemic *Helianthus exilis* and a larger recent weedy invader *Helianthus annuus* (Heiser 1949). Work using early genetic markers failed to find evidence for a hybrid origin of *H. bolanderi*, but the hybridization between *H. bolanderi* and *H. annuus* is ongoing as *H. annuus* invades California (Rieseberg *et al.* 1988; Carney *et al.* 2000). Here, we reinvestigate this classic example with high-resolution genomic data to ask whether *H. bolanderi* is a product of introgression and also whether the direction of introgression, if any, is consistent with current theory.

Correspondence: Gregory L. Owens, Fax: (1) 604 822 6089; E-mail: gregory.owens@alumni.ubc.ca

During invasion, hybridization between the invader and native species can occur and is recognized as a major issue in species conservation (Levin & Ortega 1996; Rhymer & Simberloff 1996; Vilà *et al.* 2000; Allendorf *et al.* 2001). Although contamination of the native gene pool and 'genome extinction' are the primary conservation issues, current models suggest that it is the invader that should be subject to the most introgression (Grant 1981; Currat *et al.* 2008). This is because hybrids will more often backcross with the invading species rather than the declining native species. As the invasion spreads, these backcrossed individuals will advance with the wavefront. Therefore, as the invasion continues, introgression should continue to increase until counteracted by selection. This pattern has been seen in many empirical studies (e.g. Martinsen *et al.* 2001; Donnelly *et al.* 2004; Secondi *et al.* 2006), but not all (Goodman *et al.* 1999; Carney *et al.* 2000; Takayama *et al.* 2006), and is often attributed to the effects of selection- or sex-biased dispersal (Kulikova *et al.* 2004; Melo-Ferreira *et al.* 2005).

In Californian sunflowers, contemporary hybridization with *H. annuus* appears to be limited to *H. bolanderi* and not its sister species *H. exilis*. *Helianthus annuus* is native to central USA and has invaded California from south to north, up the Central Valley over the last several thousand years (Heiser 1949). Currently, it is found primarily south of Sacramento (38.5°N) and has replaced *H. bolanderi* populations in the Central Valley over the last 100 years (Carney *et al.* 2000). Hybridization is expected to be rarer with *H. exilis* because it occurs almost exclusively on serpentine soil, an extreme soil type characterized by a high Mg/Ca ratio and high levels of heavy metals, including Ni, Cr and Cd (Brooks 1987). Serpentine soil is deadly to nonadapted plant species but is home to a wide variety of endemic species (Brady *et al.* 2005; Safford *et al.* 2005). *Helianthus bolanderi* also occurs on serpentine soil, but not exclusively, while *H. annuus* has not been reported from serpentine soils. *Helianthus exilis* is morphologically differentiated from *H. bolanderi* by having lance-linear leaves, entire leaf margins and smaller flower heads and fruit.

We used genotyping by sequencing (GBS), a popular restriction enzyme-based method for reducing genome complexity, to interrogate the genomes of these three species. We ask the following three questions. (i) Is *H. bolanderi* of hybrid origin as hypothesized by Heiser (1949) and Grant (1981)? (ii) Is there introgression between *H. bolanderi* and *H. annuus*? and (iii) Is introgression biased into the invader, *H. annuus*, as predicted by models? Our results provide the final resolution of a classic case study of the role of hybridization in plant evolution and a test of contemporary theory regarding patterns of introgression during biological invasions.

Methods

Data preparation

Sampling. We collected *Helianthus exilis* and *Helianthus bolanderi* seeds from 10 sites across the known species ranges in August 2011 (Table 1). Additionally, we used seeds from the United States Department of Agriculture

Table 1 Sample information by population. Non-*Helianthus bolanderi–exilis* samples are from a range of locations specified individually in Table S1 (Supporting information). Sample size information is after sample quality filtering

Population	Species	Sample size	Latitude	Longitude	Area	Serpentine?	Mg/Ca ratio
G100	*H. bolanderi–exilis*	10	39.40117	−122.61349	Coast Mountains	Yes	4.26
G101	*H. bolanderi–exilis*	3	39.26759	−122.48275	Coast Mountains	No	0.48
G102	*H. bolanderi–exilis*	10	39.12638	−122.43213	Coast Mountains	Yes	3.38
G103	*H. bolanderi–exilis*	10	38.7804	−122.57185	Coast Mountains	Yes	2.41
G108	*H. bolanderi–exilis*	11	38.87585	−120.8205	Sierra Nevada Mountains	Yes	2.66
G109	*H. bolanderi–exilis*	10	39.17832	−121.75977	Central Valley	No	0.16
G110	*H. bolanderi–exilis*	6	39.25156	−121.88924	Central Valley	No	0.30
G111	*H. bolanderi–exilis*	10	39.34395	−121.44869	Central Valley	No	0.14
G114	*H. bolanderi–exilis*	11	41.28199	−122.85186	North Mountains	Yes	4.53
G115	*H. bolanderi–exilis*	7	41.64306	−122.74711	North Mountains	Yes	13.02
G116	*H. bolanderi–exilis*	5	39.066322	−122.4784	Coast Mountains	Yes	NA
G118	*H. bolanderi–exilis*	9	39.2627	−122.51157	Coast Mountains	Yes	1.89
G119	*H. bolanderi–exilis*	9	39.48584	−121.31271	Sierra Nevada Mountains	No	0.26
G120	*H. bolanderi–exilis*	8	38.543	−121.7383	Central Valley	No	NA
G121	*H. bolanderi–exilis*	10	38.82395	−122.33725	Coast Mountains	Yes	NA
G122	*H. bolanderi–exilis*	8	38.73309	−122.52462	Coast Mountains	Yes	2.78
G123	*H. bolanderi–exilis*	10	39.83434	−121.58227	Sierra Nevada Mountains	Yes	6.25
G124	*H. bolanderi–exilis*	10	38.84119	−120.87647	Sierra Nevada Mountains	Yes	2.50
G127	*H. bolanderi–exilis*	10	37.84557	−120.46388	Sierra Nevada Mountains	Yes	1.82
G128	*H. bolanderi–exilis*	4	41.03086	−122.42451	North Mountains	Yes	1.85
G129	*H. bolanderi–exilis*	6	39.88756	−122.63451	Coast Mountains	No	0.84
G130	*H. bolanderi–exilis*	10	41.29794	−122.72187	North Mountains	Yes	2.56
cal_ann	*H. annuus*	24	NA	NA	California	NA	NA
cen_ann	*H. annuus*	76	NA	NA	Central USA	NA	NA
div	*H. divaricatus*	5	NA	NA	Central USA	NA	NA
gig	*H. giganteus*	5	NA	NA	Central USA	NA	NA
gro	*H. grosseserratus*	6	NA	NA	Central USA	NA	NA
max	*H. maximiliani*	10	NA	NA	Central USA	NA	NA
nut	*H. nuttallii*	3	NA	NA	Central USA	NA	NA

National Plant Germplasm System (USDA NPGS) (11 populations) and one population from Jake Schweitzer to supplement our collection (Locations plotted in Fig. 1). As there is controversy in the literature about the species' delimitation between *H. exilis* and *H. bolanderi*, we took an agnostic approach to collecting (Jain *et al.* 1992). Populations spanning the combined species ranges, including populations that had previously been identified as either species, were sampled. Similarly, all available samples of both species from the USDA NPGS were genotyped. Up to 10 seeds were sampled per population. For personally collected populations, each seed came from a separate maternal parent; for USDA NGRP seed, pooled parental seed was used. For samples from throughout the range of *Helianthus annuus* as well as for several perennial sunflower outgroup species (specifically *Helianthus divaricatus*, *Helianthus giganteus*, *Helianthus grosseserratus*, *Helianthus maximiliani* and *Helianthus nuttallii*), we employed GBS data previously generated in the Rieseberg laboratory using the same GBS protocol employed here (Baute 2015). These data are currently on the NCBI Sequence Read Archive (SRA) (Table S1, Supporting information). Altogether, we used 322 samples: 190 *H. bolanderi–exilis*, 102 *H. annuus* and 30 perennial sunflowers.

Soil sampling. For each site from which we collected seeds, we also collected soil for composition analysis. Soil was collected six inches below the surface in five randomly selected locations spanning the collection area and pooled. Soil was analysed at A&L Western Labs and measured for organic matter, phosphorous, potassium, magnesium, calcium, sulphur, pH and hydrogen. Additionally, DTPA-Sorbitol extraction was used to measure the heavy metals nickel, chromium and cobalt.

For a subset of the USDA NGRP samples, calcium and magnesium concentrations in the soil were measured (Gulya & Seiler 2002). The remaining three sites had no soil measurements, but two were from areas described as serpentine (G116, G121) and one from an area with no nearby serpentine (G120).

Genotyping by sequencing. Seeds were germinated and grown to seedling stage. DNA was extracted from young leaves using Qiagen DNeasy plant kit (Qiagen, Valencia, CA, USA), with RNase A. DNA quantity was assessed using a Qubit 2.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA).

Genotyping by sequencing library construction was done using the standard protocol of Elshire *et al.* (2011) except for the addition of a gel isolation step to eliminate dimers generated by the polymerase chain reaction (Elshire *et al.* 2011). Two libraries of 95 samples each were prepared.

Sequencing and data preparation. Both GBS libraries were paired end sequenced on an Illumina HiSeq 2000 at the UBC Biodiversity Research Center, a single lane each. Individual data were demultiplexed from within read barcodes using a custom Perl script that also removed barcode sequence. Fastq files were then trimmed for low-quality reads and Illumina adapters using Trimmomatic (Bolger *et al.* 2014). Raw demultiplexed data were uploaded to the SRA (SRP062491). All custom scripts are included in Appendix S1 (Supporting information).

SNP calling. Data were aligned to the *H. annuus* reference genome (HA412.v1.1.bronze) using BWA (version 0.7.9a) and STAMPY (version 1.0.23) using default parameters (Li & Durbin 2010; Lunter & Goodson 2011). Because we were aligning sequence data to a diverged species reference, we used STAMPY to increase alignment quality. BAM files were cleaned, sorted and had their read group information added using Picard tools (1.114) (http://broadinstitute.github.io/picard/). We used the Genome Analysis ToolKit (version 3.3) to identify possible alignment issues and realign those areas using 'RealignerTargetCreator' and 'IndelRealigner' (Van der Auwera *et al.* 2002). BAM files were processed using the GATK 'HaplotypeCaller' program and SNPs were ultimately called all together using 'GenotypeGVCFs'. SNPs were converted to a flat table format using a custom Perl script which removed indels, required sites to have QUAL > 20 and MQ > 20,and required individual genotypes to have depth between 5 and 100 000 and GT_QUAL > 20. Samples with below ~25 000 reads were removed because they did not have enough data to be informative.

After initial SNP calling, the data were divided into three data sets: only *H. bolanderi* and *H. exilis* (data set 'BE'), *H. bolanderi*, *H. exilis* and *H. annuus* (data set 'BE + A'), and all samples including the outgroup perennials (data set 'BE + A + P'). These sets were filtered to remove sites with sample coverage <60%, minor allele frequency <1% and observed heterozygosity >60% using a custom perl script. These are referred to as the 'filtered' data sets. For population structure analysis, linkage between markers can cause issues, so we subsequently thinned each filtered set so that each SNP is at least 1000 bp from its nearest neighbour, effectively picking one SNP per GBS tag. These are referred to as the 'thinned' data sets.

Evaluating the genetic structure of Helianthus bolanderi and Helianthus exilis

Population structure and admixture. To detect admixture and population structure in *H. bolanderi–exilis*, we ran

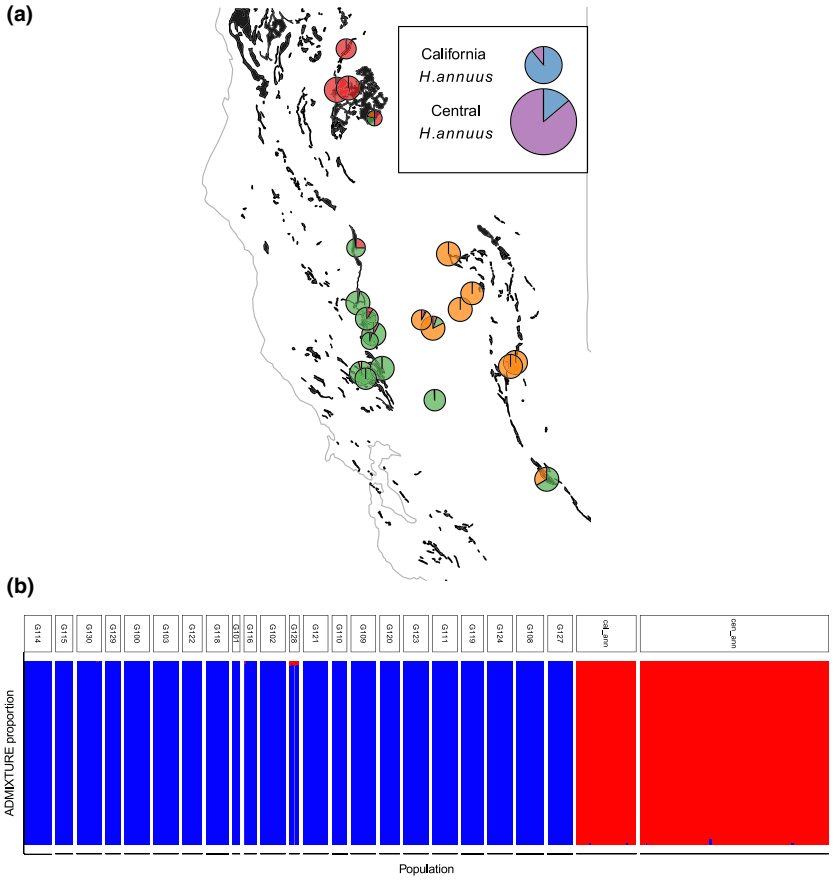

Fig. 1 (a) A map of *Helianthus bolanderi–exilis* locations with ADMIXTURE proportions (based on the filtered BE + A data set at $K = 5$) indicated by colour pie charts. Admixture group 1 (purple) and group 2 (blue) are only found in *Helianthus annuus* samples. Groups 3–5 (red, green and orange) correspond to north, west and east regions, respectively. Serpentine locations are highlighted in black on the map. (b) ADMIXTURE proportion for $K = 2$ for the filtered BE + A data set. *Helianthus bolanderi–exilis* populations are ordered by latitude. Group 1 (red) corresponds to *H. annuus* samples and group 2 (blue) to *H. bolanderi–exilis* samples.

fastStructure using the 'BE' filtered data set with $K = 1$–10 (Raj *et al.* 2014) and repeated 100 times. The optimal K was found using the 'chooseK' script bundled with fastStructure. *Admixture* was run from $K = 1$–20, using the default parameters (Alexander *et al.* 2009). Cross-validation scores were used to determine the best K value. To control for linkage effects, this was repeated with the 'thinned' data set that has neighbouring SNPs removed. Principal component analysis (PCA) was run using the 'FACTOMINER' packaged in R, using the command 'PCA'. Missing data were imputed using the package 'MISSMDA'. These analyses were repeated using the same parameters with the 'BE + A' data set.

Overall sample relatedness was visualized with an unrooted phylogenetic network using SPLITSTREE4 on the 'BE' filtered data set (Huson 1998). Uncorrected *P*-distance was used and heterozygous sites were ignored (as per defaults). This was also run using the 'BE + A + P' filtered data set.

We calculated F_{ST} between all pairs of populations using the Weir and Cockerham method (Weir & Cockerham 1984), and F_{IS} for each population (Tables S1 and S2, Supporting information). Both were calculated using custom Perl scripts.

Introgression with Helianthus annuus. To determine whether *H. bolanderi* is uniquely introgressed from *H. annuus*, we calculated Patterson's D statistic (Kulathinal *et al.* 2009; Green *et al.* 2010; Durand *et al.* 2011), which is commonly known as the ABBA–BABA test. It requires sequence data from four groups (either individual samples or allele frequencies). P1 and P2 are geographically separated populations of one species, P3 is a separate species in sympatry with P2, and P4 is an outgroup species. The test counts the number of ABBAs (where P2 and P3 share a derived allele) and BABAs (where P1 and P3 share a derived allele). Under incomplete lineage sorting, we would expect an equal number of ABBAs and BABAs, but if there is gene flow between P2 and P3, there will be excess ABBAs and D will be positive.

Since we had many samples of each group, we used allele frequencies instead of instance counts of single samples (Martin *et al.* 2015). The four groups used were all central *H. annuus* (i.e. all *H. annuus* not in California), all California *H. annuus*, an *H. bolanderi–exilis* population and all perennial sunflowers. Perennial sunflowers included *H. maximiliani*, *H. nuttallii*, *H. divaricatus*, *H. giganteus* and *H. grosseserratus*. This

monophyletic group of species is an outgroup to the annual sunflowers that include *H. annuus* and *H. bolanderi–exilis*. Only biallelic sites for which all perennial samples were fixed for a single allele were used, because these sites gave the most confidence in determining the ancestral allele. We also calculated f_d, a measure of the amount of the genome involved in introgression (Martin *et al.* 2015). For each statistic, we calculated standard deviation, *Z*-score and *P*-value using a block jackknife approach with 10 Mb size blocks (Green *et al.* 2010). This test was run on each individual *H. bolanderi–exilis* population as well as all *H. bolanderi–exilis* samples together.

For this test, a positive D score indicates that ABBA > BABA, and California *H. annuus* and *H. bolanderi–exilis* share more derived alleles. A negative D score indicates that BABA > ABBA and central *H. annuus* and *H. bolanderi–exilis* share more derived alleles. The neutral expectation under no gene flow is ABBA = BABA and D = 0.

To evaluate hypotheses about introgression, we plotted D and f_d in 10 Mb windows across the genome. We also used the *H. annuus* genetic map to compare recombination rate and introgression in 10 Mb windows using a type III ANOVA (Renaut *et al.* 2013).

A positive D statistic using allele frequencies from all samples may be driven by a subset of samples if introgression is not uniform among California *H. annuus* and *H. bolanderi–exilis* samples. It could also be caused by unmeasured introgression into central *H. annuus* by a third species [e.g. *Helianthus petiolaris*, which is known to hybridize and is largely sympatric across the central USA range of *H. annuus* (Yatabe *et al.* 2007)]. To account for this, we used a subsampling strategy that isolates each sample individually (while retaining all samples for other groups) and calculates a D score. For example, one test would include one central *H. annuus* sample, all Californian *H. annuus*, all *H. exilis–bolanderi* and all perennial samples. Thus, for each sample we get a D score reflecting its effect on the overall D score. Significance was calculated using a block jackknife approach (as above).

We use these single sample D scores to assess the hybrid origin of *H. bolanderi*. If *H. bolanderi* was a hybrid species, we would expect all *H. bolanderi–exilis* samples to have to fall into two distinct sets: one with high D scores (representing the hybrid *H. bolanderi*) and one with lower, but possibly still positive, D scores (representing nonintrogressed *H. exilis*). A nonintrogressed *H. exilis* may still produce a positive D score because of introgression in *H. annuus*, but a hybrid species should be distinctly higher.

To evaluate the amount of introgression in each sample or population, we plotted individual sample D scores vs. latitude (for *H. bolanderi–exilis* and *H. annuus*) and vs. collection date (for *H. annuus*) (Wickham 2009). We used a type III ANOVA, using the R package 'CAR', to determine whether each of these factors affected D or f_d (R Development Core Team 2008; Fox & Weisberg 2010).

Testing the directionality of gene flow with Helianthus annuus

The partition D test. A positive D score indicates gene flow, but does not specify if the gene flow is into *H. bolanderi–exilis*, into *H. annuus*, or is bidirectional. To answer this question, we used the partitioned D statistic (Eaton & Ree 2013). This extension of the ABBA–BABA test uses five taxa instead of four and can determine directionality of introgression using a set of three different tests. The main difference between the partitioned D statistic and Patterson's D statistic is that the partitioned version divides the P3 clade (i.e. *H. bolanderi–exilis* in our analysis) into two lineages, $P3_1$ and $P3_2$, which are assumed not to be exchanging genes. The three partitioned D statistic tests then ask whether the enrichment of shared derived alleles shown by the positive classic D statistic is from the first, second or both P3 lineages. Specifically, D_1 compares counts of ABBAA and BABAA looking for enriched shared derived alleles specifically in $P3_1$, D_2 compares counts of ABABA and BAABA looking for enriched shared derived alleles specifically in $P3_2$, and D_{12} compared counts of ABBBA and BABBA looking for enriched shared derived alleles in both $P3_1$ and $P3_2$.

Comparing the results of the three tests can be used to determine the directionality of gene flow. Consider the scenario where D_{12} is positive. This either suggests gene flow from P2 into the ancestor of $P3_1$ and $P3_2$, gene flow from P2 into both $P3_1$ and $P3_2$, or gene flow from $P3_x$ into P2. If the first two scenarios can be ruled out by other tests or outside information, then gene flow in one direction is supported. In this scenario, the lineage of P3 that is donating genes is determined by the D_1 and D_2 tests. This in itself only indicates that gene flow is going in at least one direction, not that it is unidirectional, but by rotating the positions in the phylogeny (i.e. P1→$P3_2$, P2→$P3_1$, $P3_1$→P2, $P3_2$→P1), and repeating the tests we can make a case for the overall directionality of gene flow. For example, if in the rotated phylogeny scenario the D_{12} test is zero, then there is a lack of evidence for gene flow in the opposite direction and unidirectional gene flow is supported overall. With this framework in mind, we used two phylogenetic scenarios (i.e. the same phylogeny rotated differently) to get at the directionality of gene flow.

The first scenario uses the five groups in the following order: P1 = all central *H. annuus*, P2 = all California *H. annuus*, $P3_1$ = a southern *H. bolanderi–exilis* population, $P3_2$ a northern *H. bolanderi–exilis* population (G115) and P4 = perennial outgroup. In this case, we are treating G115 as nonintrogressed due to its geographic isolation from any *H. annuus* population and the strong population structure, indicating little within species gene flow.

With our groupings in mind, the three tests from the partitioned D have different implications in this scenario. D_{12} asks whether derived alleles found in both *H. bolanderi–exilis* populations are more often found in California *H. annuus*, than central *H. annuus*. A positive score suggests gene flow from any *H. bolanderi–exilis* into *H. annuus* because otherwise the derived allele would not be present in both *H. bolanderi* and *H. exilis* populations. D_1 asks whether derived alleles, not found in northern *H. bolanderi–exilis*, are present in California *H. annuus*. A positive score suggests that there is gene flow between the southern *H. bolanderi–exilis* and California *H. annuus*, or that there is gene flow between California *H. annuus* and a population of *H. bolanderi–exilis* more closely related to the southern *H. bolanderi–exilis* population tested. D_2 asks the same as D_1 but with northern and southern *H. bolanderi–exilis* populations reversed (i.e. this may suggest gene flow with northern *H. bolanderi–exilis* or close relative).

The test was repeated using each *H. bolanderi–exilis* population in $P3_1$, except G115, which is always in $P3_2$. This means that we did each test 21 times and our main reported result is how many of these tests were significantly positive. The number of positive tests is indicative of how consistent the signal is across the range of *H. exilis–bolanderi*. Since we tested every population, some tests involve two *H. bolanderi–exilis* populations that are both in the northern clade.

The second scenario involves a rotated phylogeny. The five groups are as follows: P1 = a northern *H. bolanderi–exilis* (G115), P2 = a southern *H. bolanderi–exilis*, $P3_1$ = California *H. annuus*, $P3_2$ = central *H. annuus* and P4 = perennial outgroup. In this scenario, D_{12} asks whether derived alleles found in all *H. annuus* are present in the southern *H. bolanderi–exilis* and not the northern. A positive score indicates gene flow into *H. bolanderi–exilis*. Tests D_1 and D_2 ask whether there are an excess of derived alleles from California *H. annuus* or central *H. annuus*, respectively, in southern *H. bolanderi–exilis*. Similarly in this scenario, we also repeat each test using a different southern *H. bolanderi–exilis* population and report the number of significantly positive tests.

For these tests, we used allele frequencies instead of individual genomes and only included sites where all perennial samples were fixed for a single allele. Significance was tested using block jackknife bootstrapping, as before, and $P < 0.05$ was used as the *P*-value cut-off. All tests were repeated using another population (G114) as the northern nonintrogressed *H. bolanderi–exilis* population.

Demographic modelling. To explore the amount and direction of gene flow, we simulated the demographic history using δaδi (Gutenkunst *et al*. 2009). δaδi simulates the site frequency spectrum of demographic scenarios and uses diffusion approximation to explore the parameter space. In our model, we use three populations (*H. bolanderi–exilis*, central *H. annuus* and California *H. annuus*) and seven parameters: three effective population sizes, N_{BE}, N_{CenA} and N_{CalA}; two times, T_1 and T_2; and two migration rates, $m_{CalA \to BE}$ and $m_{BE \to CalA}$. At time T_1, central *H. annuus* and *H. bolanderi–exilis* diverge, and at time T_2, *H. annuus* invades California and exchanges genes with *H. bolanderi–exilis* until present (Fig. 2). We also ran the model with the migration events removed in all combinations.

We used the Broyden–Fletcher–Goldfarb–Shanno optimization method to fit parameters for each model. Searches were started from 10 randomly perturbed starting positions with up to five iterations each. The best-fit parameters were used for a further optimization for up to 20 iterations. Samples were extrapolated to grid size of (175, 75, 25) to maximize the number of usable SNPs. Three hundred bootstrap site frequency spectra were generated using 1 Mb block bootstrapping. This was used to calculate confidence intervals for all parameters. Parameters were corrected using the mutation rate of 6.1×10^{-9} substitutions/site/generation (Sambatti *et al*. 2012). Effective sequenced length was estimated by measuring the number of sites with >5 reads in 88 *H. bolanderi–exilis*, 38 central *H. annuus* and 13 California *H. annuus* samples, including invariant sites. These numbers were chosen to reflect the extrapolation grid size.

Results

Sample and SNP information

Sample sizes. We removed three *Helianthus bolanderi–exilis* and two *Helianthus annuus* samples for having <25 000 reads. One perennial sample (GB148) was removed because it grouped with *H. annuus* samples in the splits network analysis. After removing samples, we had sequence data for 187 *H. bolanderi–exilis* samples,

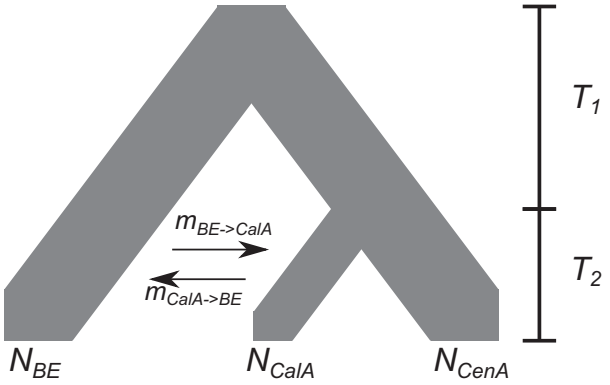

Fig. 2 Demographic scenario modelled in δaδi, including all modelled parameters. Including effective population size (*N*) for *Helianthus bolanderi–exilis* ($_{BE}$), California *Helianthus annuus* ($_{CalA}$), central *H. annuus* ($_{CenA}$), migration rates (*m*) and time (*T*).

100 *H. annuus* samples and 29 perennial sunflower samples (Table S1, Supporting information).

Soil analysis. Serpentine sites are primarily characterized by Mg/Ca ratio > 1 (Jenny 1980). All sites identified by plant composition and soil maps as serpentine were confirmed with soil measurements (Table S1, Supporting information).

SNP calling. All demultiplexed data were uploaded to the SRA (SRP062491). Number of reads per sample and per cent aligned reads are listed in Table S1 (Supporting information). After initial filtering for quality and depth, we found 131 150 SNPs total (Table 2). Subsequent filtering for coverage (>60%), minor allele frequency (>1%) and observed heterozygosity (<60%) reduced that to 9593 SNPs.

Population structure and introgression with Helianthus annuus

Population structure approaches. ADMIXTURE and fastStructure suggest a fractal pattern of divergence in *H. bolanderi–exilis* based on geography rather than soil type. At $K = 2$, east and west populations are separated; at $K = 3$, northern populations become their own group; and at $K = 4$, south-west populations separate (Figs S1 and S2, Supporting information). At higher *K* values, individual populations become their own group and intermediate or admixed individuals are rare (Figs S1 and S2, Supporting information). Both ADMIXTURE and fastStructure generally agree on cluster assignment for lower *K* values (2–4) but above that there is inconsistency between runs and methods.

Substantial admixture between *H. annuus* and *H. bolanderi–exilis* was not seen in either ADMIXTURE or fastStructure results (Figs 1, S1 and S2, Supporting information). At $K = 2$, *H. annuus* and *H. bolanderi–exilis* are separate groups with the possible exception of the *H. bolanderi–exilis* population G128. ADMIXTURE showed G128 to have 1–2% ancestry from the *H. annuus* group. In fastStructure, this population had slightly elevated *H. annuus* ancestry but of a lower magnitude (~0.5% admixed ancestry).

Table 2 Number of SNPs found for each data set. The filtered data set removed sites where sample coverage < 60%, observed heterozygosity > 60% or minor allele frequency < 1%. The thinned data set reduced the filtered data set down to one SNP per 1000 bp

	Total variant sites	Filtered	Thinned
Only *Helianthus bolanderi–exilis* 'BE'	57 926	7514	1183
H. bolanderi–exilis and *Helianthus annuus* 'BE + A'	103 318	8915	1095
All samples 'BE + A + P'	131 150	9593	1062

SPLITSTREE4 and PCA recapitulated the results seen in ADMIXTURE and fastStructure (Figs 3 and 4). For the splits network, *H. bolanderi–exilis*, *H. annuus* and the perennial species form monophyletic groups without admixture. In the PCA, the first principal component separated *H. annuus* and *H. bolanderi–exilis*, and the second separated the east and west *H. bolanderi–exilis* populations.

ADMIXTURE cross-validation testing found $K = 8$ for BE and $K = 6$ for BE + A to have the lowest error, although scores were relatively flat from $K = 5$–10 (Fig. S3, Supporting information). For fastStructure, marginal likelihood was universally maximized at $K = 2$ for BE and $K = 3$ for BE + A. The *K* value that best explained population structure depended on the run and data set: BE filtered = 3–5, BE thinned = 3–7, BE + A filtered = 3–4, BE + A thinned = 3–6. We do not further evaluate the best *K* value beyond the fact that *H. bolanderi–exilis* and *H. annuus* are never placed in the same group and that there is some level of geographic structure in *H. bolanderi–exilis*. The exact best *K* value to explain the geographic structure is not relevant to our hypotheses.

F_{ST} values between populations of *H. bolanderi–exilis* were high (0.041–0.509, mean = 0.331), implying minimal gene flow between geographically distant populations (Table S2, Supporting information). Between *H. bolanderi–exilis* and *H. annuus*, F_{ST} was also very high (mean $F_{ST} = 0.508$ and 0.472 for Californian and central *H. annuus*, respectively).

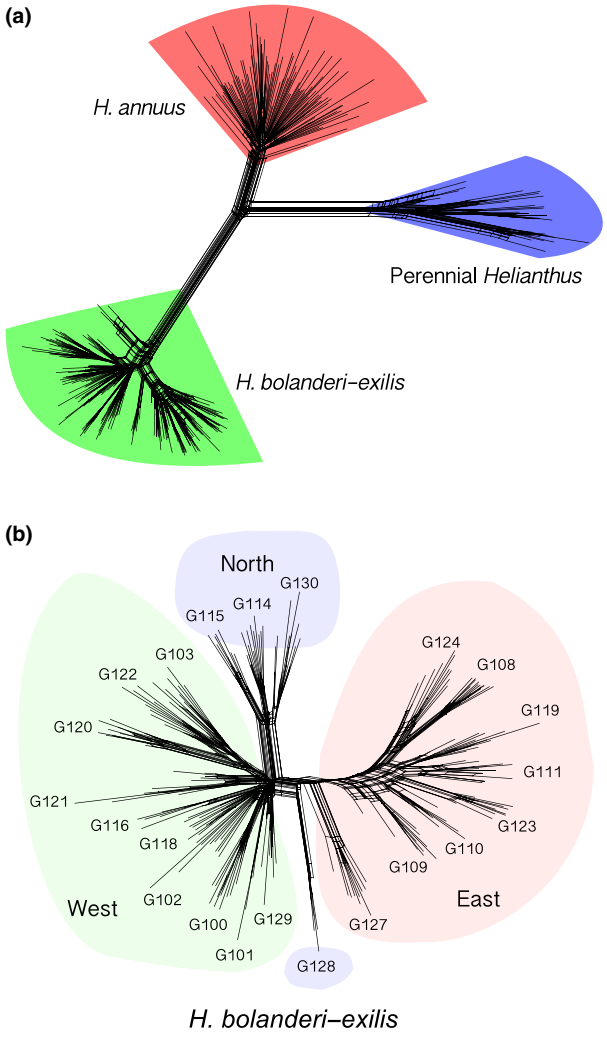

Fig. 3 Splits network analysis of (a) the filtered BE + A + P data set and (b) the filtered BE data set. Network was made using SPLITSTREE4 with uncorrected *P*-distance.

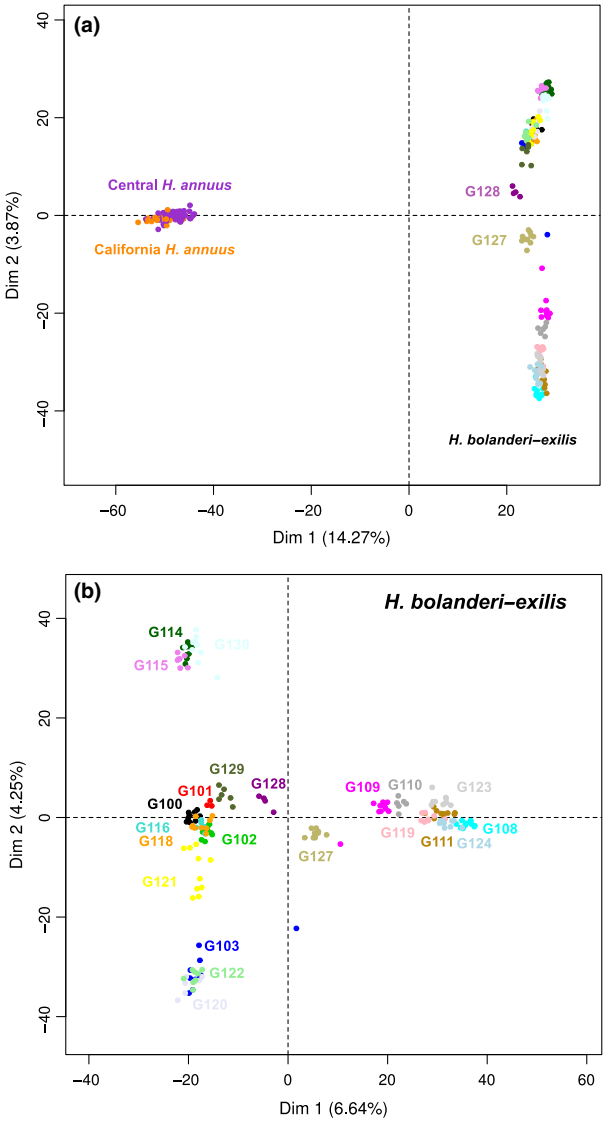

Fig. 4 Principal component analysis of (a) the filtered BE + A data set and (b) the filtered BE data set. In (a) populations G127 and G128 are labelled because they occupy the most intermediate position in the *Helianthus bolanderi–exilis* cluster.

F_{IS} showed no evidence of inbreeding in *H. bolanderi–exilis* populations, consistent with their self-incompatibility (Table S1, Supporting information). Moderate inbreeding was observed in *H. annuus* and several perennial species, likely because samples from multiple populations were pooled and any population structure will result in increased F_{IS} (Wahlund 1928).

ABBA–BABA tests. We found a significant positive D score (suggesting Californian *H. annuus–H. bolanderi–exilis* gene flow) for the full data set (0.123 ± 0.033, P = 1.6e-4) and for all individual *H. bolanderi–exilis* populations (Fig. 5a). The fraction of the genome shared through introgression was overall 5–8% (f_d = 0.065 ± 0.017). When visualized across the genome, the amount of introgression was variable (Fig. S4, Supporting information). In particular, chromosome Ha1 had high amounts of introgression, while introgression was low on Ha2, Ha11, Ha12 and Ha15 (Table S3, Supporting information). When D or f_d is compared with recombination rate in *H. annuus*, there is no association ($P > 0.1$) (Fig. S5, Supporting information).

When looking at the effect of individual samples, we find positive D scores with 70/76 central *H. annuus* samples, 21/24 California *H. annuus* and 187/187 *H. bolanderi–exilis* samples (Table S3, Fig. S6, Supporting information). Population G128, which exhibited slight evidence of admixture in the ADMIXTURE analysis, showed slightly below average D scores. We find no relationship between collection date or latitude and D

or f_d for the California *H. annuus* samples (all $P > 0.12$) (Fig. S7, Supporting information), but latitude does correlate with D and f_d in *H. bolanderi–exilis* (D: $F_{1,183} = 24.0$, $P <$ e-5; f_d: $F_{1,183} = 17.3$, $P <$ e-4) (Fig. S8, Supporting information).

Directionality of gene flow with Helianthus annuus

Partitioned D tests. The partitioned D statistic using scenario 1 produced D_{12}, D_1 and D_2 tests that were significantly positive for 21/21, 17/21 and 0/21 populations, respectively. For scenario 2, the number of significantly positive populations was 0/21, 2/21 and 0/21, respectively (Fig. 5b). In scenario 2, test D_2, three populations produced significantly negative values (Table S4, Supporting information). Using G114 as the reference northern population produced similar results (Table S4, Supporting information).

Demographic modelling. Demographic modelling found the most likely model included bidirectional gene flow (Table 3). Both the unidirectional gene flow models were better than no migration (into California *H. annuus*: $P = 0.0012$; into *H. bolanderi–exilis*: $P = 0.0059$). Bidirectional gene flow was better supported than either unidirectional model (into California *H. annuus*: $P = 0.0055$; into *H. bolanderi–exilis*: $P = 0.0046$).

In the best-supported model, effective population size of central *H. annuus* is ~880 000, of California *H. annuus* is ~95 000 and of *H. bolanderi–exilis* is ~490 000. The model estimated ~410 000 years ago for the *H. annuus–H. bolanderi–exilis* split and 18 000 years ago for when *H. annuus* invaded California. Migration rates were below one migrant per generation (between 0.08 and 0.5).

Discussion

The nonhybrid origin of Helianthus bolanderi

Using our high-resolution genomic data, we can definitively rule out the putative hybrid origin theory of *Helianthus bolanderi*, confirming early work by Rieseberg *et al*. (1988). Principal component, population structure and phylogenetic network analysis all fail to find evidence for admixture between a subset of *H. bolanderi–exilis* and *Helianthus annuus*. If *H. bolanderi* were of hybrid origin, we would expect some of our sampled populations (particularly those in the eastern part of the range where *Helianthus exilis* is not present) to be genetically closer to *H. annuus*, but we do not see this. This does not mean that there is no gene flow with *H. annuus* and, indeed, our ABBA–BABA testing shows that there is.

As a secondary hypothesis, we evaluated the possibility that *H. bolanderi* had undergone greater introgression with *H. annuus* than did *H. exilis*. The phenotypic intermediacy that motivated the hybrid origin hypothesis might be caused by small amounts of introgression, less than what is typically envisioned for a hybrid species, and this may not be detected by the coarser population structure or clustering analyses. However, using the ABBA–BABA test, we failed to find support for this possibility as well. All *H. bolanderi–exilis* popula-

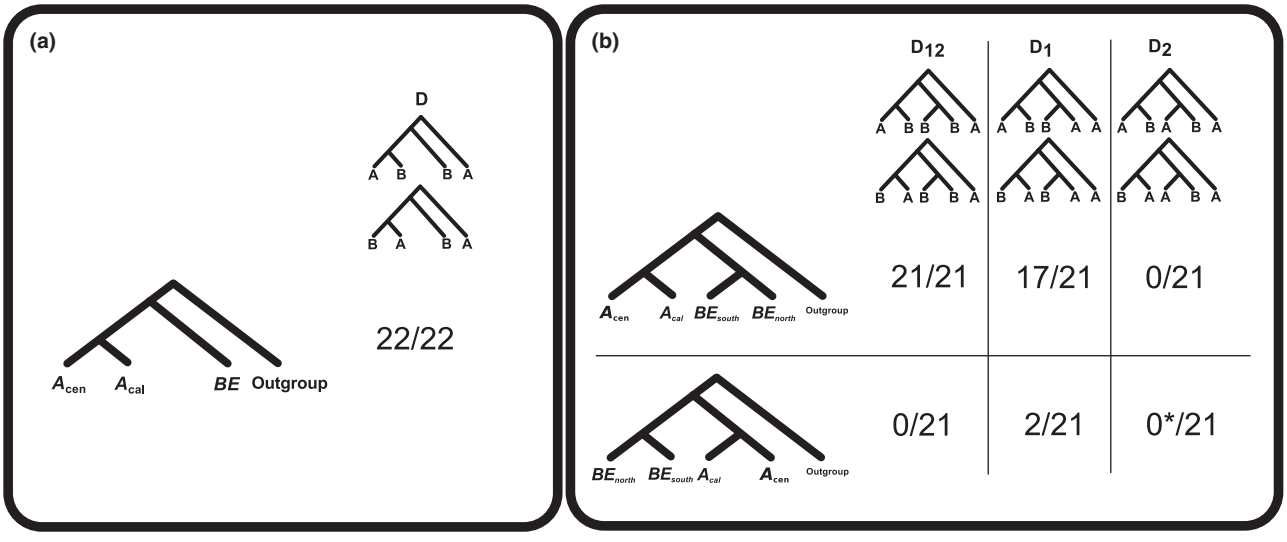

Fig. 5 Number of significantly positive tests using (a) the Patterson's D statistic and (b) the partitioned D statistic. (a) Each test uses a separate *Helianthus bolanderi–exilis* population. (b) Each test uses a different *H. bolanderi–exilis* population in the BE_{south} position but keeps BE_{north} constant as G115. Phylogenetic scenarios being compared are included in each test diagram.

Table 3 Parameters for all δaδi models. Confidence intervals based on block bootstrapping. Migration is scaled to the number of migrants per generation in the receiving population

	No migration		Into BE migration		Into CalA migration		Bidirectional migration	
	ML	95% CI	ML	95% CI	ML	95% CI	ML	95% CI
LL	−7494.10	—	−6605.07	—	−7262.47	—	−6464.80	—
θ	469.66	—	321.84	—	321.85	—	313.91	—
N_{BE} ($\times 10^5$)	5.70	5.65–5.75	4.96	4.85–5.07	4.05	4–4.09	4.94	4.83–5.05
N_{CenA} ($\times 10^5$)	8.46	8.26–8.65	8.77	8.55–8.99	6.07	5.93–6.22	8.80	8.58–9.02
N_{CalA} ($\times 10^5$)	0.97	0.87–1.07	1.21	1.21–1.21	0.49	0.48–0.5	0.95	0.94–0.95
T_1 ($\times 10^5$)	3.15	3.12–3.18	3.97	3.88–4.06	2.36	2.34–2.39	4.14	4.07–4.22
T_2 ($\times 10^5$)	0.19	0.17–0.21	0.22	0.22–0.22	0.10	0.1–0.1	0.18	0.18–0.18
$m_{CalA \to BE}$	—	—	0.45	0.44–0.46	—	—	0.48	0.47–0.5
$m_{BE \to CalA}$	—	—	—	—	0.11	0.06–0.17	0.08	0.05–0.11

tions show positive D scores—there is no bimodality that can be attributed to two species, one of which hybridizes (although northern populations show some reduction in D, discussed below) (Fig. S7, Supporting information). In fact, our results do not support *H. exilis* and *H. bolanderi* as separate species, but are more consistent with a single species with population structure associated with geographic location. The division between *H. exilis* and *H. bolanderi* has been a point of contention in the literature. Originally (and currently) designated as different species, they have also been classified as two subspecies, and two species plus one ecotype (Grey 1865; Heiser 1949; Jain *et al.* 1992). Further complicating this, the currently recognized morphological differences between the species, leaf shape, flower head size and seed size can be confounded by phenotypic plasticity and the stunting effect of serpentine soil making in situ species identification difficult. Herbarium records for both species suggest that *H. exilis* is found in the North Coast and Klamath Ranges of California, while *H. bolanderi* entirely encompasses that range and extends south and east into the northern Central Valley and Sierra Nevada foothills. Our genetic data tell a different story.

At the highest level, populations are divided into east and west clades. Although this roughly corresponds to the ranges of *H. bolanderi* and *H. exilis,* respectively, both clades are not present in the western range as expected based on current descriptions of species' ranges. Furthermore, the next level of population structure separates the northern populations from the rest, again inconsistent with two overlapping species. F_{ST} between populations is quite high, even for populations relatively close together and all populations are monophyletic within the splits network analysis.

Taken together, this suggests a single species with many isolated populations. Future work should assess phenotypic variation in a common garden and hybrid sterility for crosses between samples in the eastern, western and northern clades to determine whether they are reproductively isolated. It could also establish whether the phenotypic differences purported between *H. exilis* and *H. bolanderi* follow the genetic divides we show here. We tentatively call the combined species, *H. bolanderi*. Both species names were published in the same issue by Asa Grey in 1865, but *H. bolanderi* was listed first and was considered to be the more widespread species (Grey 1865).

Gene flow with Helianthus annuus

The genetic data we present here show evidence for introgression between *H. annuus* and *H. bolanderi–exilis.* Although both population structure and clustering analyses do not show signs of admixture, the Patterson's D statistic is clear that introgression has occurred in California. When testing the effect of individual samples, we found the vast majority produced positive D scores (Fig. S6, Supporting information). This shows that the signal we are seeing is not from ghost introgression in a minority of samples (i.e. the effect of *H. petiolaris* introgression in central *H. annuus*). What the overall D statistic does not tell us is which way gene flow is occurring (e.g. *H. bolanderi–exilis* into *H. annuus*, *H. annuus* into *H. bolanderi–exilis* or bidirectional). To get at the direction of introgression, we used the partitioned D statistic with two phylogenetic scenarios (Eaton and Ree 2013). In both of these, we treat the most northern *H. bolanderi–exilis* population as nonintrogressed. We make this assumption for two reasons: (i) *H. annuus* is largely limited to the southern half of California and excluded from serpentine regions. The most northern *H. bolanderi–exilis* population (G115) is deep in a Klamath Mountains, far from the range of *H. annuus* and on a serpentine patch. (ii) The high population structure and isolated nature of populations in *H. bolanderi–exilis*

means that gene flow is low between populations and unlikely to have spread introgressed alleles that far in the relatively short period of time that *H. annuus* has been in California.

The partitioned D statistics show that gene flow is largely from *H. bolanderi–exilis* into *H. annuus*. This is seen critically in test D_{12} in both scenarios (Fig. 5). For scenario 1, D_{12} shows that derived alleles present in both *H. bolanderi–exilis* populations are enriched in the California *H. annuus* samples. This must be because of gene flow into *H. annuus* from *H. bolanderi–exilis* because the reverse could not spread the alleles to both populations. One alternative scenario is that gene flow occurred before the *H. bolanderi–exilis* populations diverged, but considering the high F_{ST} between populations of *H. bolanderi–exilis* and recent invasion of California by *H. annuus*, it is highly improbable that *H. annuus* was in California before *H. bolanderi–exilis* spread to its current range. For scenario 2, D_{12} is never significant. This shows that the southern populations are not enriched for derived alleles present in all *H. annuus* populations, as would be expected whether gene flow was bidirectional. Together, these results suggest unidirectional gene flow from *H. bolanderi–exilis* into *H. annuus*. The other tests of the partitioned D statistic (D_1 and D_2) also agree with this interpretation and dissected fully in Appendix S2 (Supporting information).

Demographic modelling supports bidirectional gene flow in California (Table 3). This is in partial conflict with the partitioned D statistic results. These methods use different ways of detecting gene flow; δaδi models demographic scenarios that produce similar site frequency spectra to the empirical data while the partitioned D statistic looks for imbalances in inheritance scenarios within a phylogeny. δaδi would not actually use information about shared derived alleles that is driving the partitioned D statistic signal. It is also possible that demographic modelling is affected by the population structure within the *H. bolanderi–exilis* samples. On the other hand, the partitioned D statistic may be underpowered for some scenarios and gene flow may be bidirectional, but unequal (i.e. there is gene flow into *H. bolanderi–exilis* but not enough to detect). Thus, we have conclusive evidence of gene flow into California *H. annuus* and ambiguous signals of the reverse; therefore, gene flow appears to be stronger into California *H. annuus*.

Theory by Currat *et al.* (2008) predicts that in this scenario the invader should have more introgressed alleles than the native species. Our results provide support for this theory—introgression does appear to be stronger into the invader *H. annuus*. Although we might expect introgression to be greater in more northern *H. annuus* populations (since they are in greater contact with *H. bolanderi–exilis*) or in populations collected at a later year (if introgression is ongoing), D scores for individual samples are not correlated with latitude or collection date (Fig. S8, Supporting information). This is also counter to theory that predicts greater introgression in populations on the range edge (i.e. northern samples). This counter-intuitive result may be because the spread of *H. annuus* across California was not a simple expanding wave and hybridization occurred haphazardly or that hybridization occurred late in expansion and only some lineages were affected. Furthermore, the model used by Currat *et al.* does not include reproductive isolation between the species and there is a significant sterility barrier between *H. bolanderi–exilis* and *H. annuus* (Chandler *et al.* 1986).

The Patterson's D statistic is positive in all *H. bolanderi–exilis* populations, but has regional variation. Specifically, the four northern populations have lower D statistics than the rest (mean 0.126 vs. 0.187, students *t*-test $P <$ e-13). This may be due to introgression in southern and central populations or, more likely, that introgressed alleles in *H. annuus* came from more southerly populations. The amount of introgression is not evenly spread across the genome; several chromosomes do not show evidence of introgression, in particular Ha2, Ha11, Ha12 and Ha15 (Fig. S4, Table S3, Supporting information). Previous work has shown associations between low recombination rate and reduced introgression, but we do not see that in our data (Barton 1979; Machado *et al.* 2007; Yatabe *et al.* 2007; Fig. S5, Supporting information). This may be because we do not have a genetic map of *H. bolanderi–exilis*, so our estimates of recombination rate are missing the major effects of chromosomal rearrangements. Chromosomal rearrangements are known to reduce introgression in sunflowers and other species (White 1978; Rieseberg 2001; Giménez *et al.* 2013; Barb *et al.* 2014) and, indeed, pollen sterility and meiotic abnormalities indicate there are several rearrangements between *H. annuus* and *H. bolanderi–exilis* (Chandler *et al.* 1986). Particularly, high values of introgression are seen in Ha1, perhaps from positive selection on loci or more neutrally from allele surfing (Hallatschek & Nelson 2008). Alternatively, simulation studies have shown that localized high D values may be due to the reduced D_{xy} in the absence of gene flow so variation in D may be a side effect of this and not reflect true gene flow variation (Martin *et al.* 2015).

Edaphic quality and introgression

The toxicity of serpentine soil excludes *H. annuus* migrants. Consequently, we would expect to see greater

introgression in nonserpentine populations of *H. bolanderi–exilis* because both species can coexist off serpentine sites. In our data, this is not the case, and Patterson's D scores of nonserpentine samples are not significantly lower than serpentine samples (Student's *t*-test, $P = 0.1097$). This is consistent with our hypothesis that the samples we sequenced of *H. bolanderi–exilis* are not actually introgressed. Despite this, the hybridization between *H. bolanderi–exilis* and *H. annuus* most likely occurred on nonserpentine soil in California's Central Valley. Populations within the southern extent of this area collected in the 1950s are no longer present possibly due to genetic swamping by *H. annuus*. Extant nonserpentine samples appear to be in danger of a similar fate as *H. annuus* spreads north.

Conclusion

The classic example of *Helianthus bolanderi–Helianthus annuus* introgression is incorrect on several fronts: (i) *H. bolanderi* is not of hybrid origin; (ii) *H. bolanderi* cannot be distinguished genetically from one of its putative parents, *Helianthus exilis*; and (iii) the intermediate phenotype of Central Valley populations of *H. bolanderi* may be a product of parallel adaptation to an environment similar to that preferred by *H. annuus*. However, *H. annuus* has absorbed alleles from *H. bolanderi–exilis* during its invasion, but whether this introgression has any adaptive significance is unclear. The greatest test of adaptive introgression will be whether *H. annuus* can spread onto serpentine soil. Currently, it is not found on serpentine areas, but it has spread to regions neighbouring serpentine *H. bolanderi–exilis* populations. Future work could track its spread to see whether adaptive introgression of serpentine tolerance occurs.

Acknowledgements

The authors would like to thank Dan Bock for perennial sunflower sequence data, Kieran Samuk and Diana Rennison for comments on the manuscript, Brook Moyers, Kieran Samuk, Kate Ostevik, Jake Schweitzer, Tom Gulya, Laura Marek and the Donald and Sylvia McLaughlin Reserve for collection assistance, and the USDA GRIN for seed. This work was partially supported by NSERC CGS-D (GLO) and NSERC discovery grant (LHR) (Grant ID 327426).

References

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655–1664.

Allendorf FW, Leary RF, Spruell P, Wenburg JK (2001) The problems with hybrids: setting conservation guidelines. *Trends in Ecology and Evolution*, **16**, 613–622.

Barb JG, Bowers JE, Renaut S *et al.* (2014) Chromosomal evolution and patterns of introgression in *Helianthus*. *Genetics*, **197**, 969–979.

Barton NH (1979) Gene flow past a cline. *Heredity*, **43**, 333–339.

Baute GJ (2015) *Genomics of sunflower improvement from wild relatives to a global oil seed*. PhD Dissertation, University of British Columbia, Vancouver.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Brady KU, Kruckeberg AR, Bradshaw H Jr (2005) Evolutionary ecology of plant adaptation to serpentine soils. *Annual Review of Ecology, Evolution, and Systematics*, **36**, 243–266.

Brooks RR (1987) *Serpentine and Its Vegetation: A Multidisciplinary Approach*. Dioscorides Press, Kent.

Carney S, Gardner K, Rieseberg L (2000) Evolutionary changes over the fifty-year history of a hybrid population of sunflowers (*Helianthus*). *Evolution*, **54**, 462–474.

Chandler JM, Jan CC, Beard BH (1986) Chromosomal differentiation among the annual *Helianthus* species. *Systematic Botany*, **11**, 354–371.

Currat M, Ruedi M, Petit RJ, Excoffier L (2008) The hidden side of invasions: massive introgression by local genes. *Evolution*, **62**, 1908–1920.

Donnelly MJ, Pinto J, Girod R, Besansky NJ, Lehmann T (2004) Revisiting the role of introgression vs shared ancestral polymorphisms as key processes shaping genetic diversity in the recently separated sibling species of the *Anopheles gambiae* complex. *Heredity*, **92**, 61–68.

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

Eaton DAR, Ree RH (2013) Inferring phylogeny and introgression using RADseq data: an example from flowering plants (Pedicularis: Orobanchaceae). *Systematic Biology*, **62**, 689–706.

Elshire RJ, Glaubitz JC, Sun Q *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, **6**, e19379.

Fox J, Weisberg S (2010) *An R Companion to Applied Regression*. Sage publications, Thousand Oaks, California.

Giménez MD, White TA, Hauffe HC, Panithanarak T, Searle JB (2013) Understanding the basis of diminished gene flow between hybridizing chromosome races of the house mouse. *Evolution*, **67**, 1446–1462.

Goodman SJ, Barton NH, Swanson G, Abernethy K, Pemberton JM (1999) Introgression through rare hybridization: a genetic study of a hybrid zone between red and sika deer (genus *Cervus*) in Argyll, Scotland. *Genetics*, **152**, 355–371.

Grant V (1981) *Plant Speciation*. Columbia University Press, New York.

Green RE, Krause J, Briggs AW *et al.* (2010) A draft sequence of the Neandertal genome. *Science*, **328**, 710–722.

Grey A (1865) *Helianthus bolanderi*. *Proceedings of the American Academy of Arts and Sciences*, **6**, 544–545.

Gulya TJ, Seiler GJ (2002) *Plant Exploration Report*. USDA Northern Crop Science Laboratory, Fargo, North Dakota.

Gutenkunst RN, Hernandez RD, Williams SH, Bustamante CD (2009) Inferring the joint demographic history of multiple

populations from multidimensional SNP frequency data. *PLoS Genetics*, **5**, e1000695.
Hallatschek O, Nelson DR (2008) Gene surfing in expanding populations. *Theoretical Population Biology*, **73**, 158–170.
Heiser CB (1949) Study in the evolution of the sunflower species *Helianthus annuus* and *H. bolanderi*. *University of California Press*, **12**, 157–196.
Huson DH (1998) SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, **14**, 68–73.
Jain SK, Kesseli R, Olivieri A (1992) Biosystematic status of the serpentine sunflower, *Helianthus exilis* Gray. In: *The Vegetation of Ultramafic (Serpentine) Soils* (eds Proctor J, Baker AJM, Reeves RD), pp. 391–408. Intercept Limited, Andover, UK.
Jenny H (1980) *The Soil Resource*. Ecological Studies, New York, New York.
Kulathinal RJ, Stevison LS, Noor MAF (2009) The genomics of speciation in *Drosophila*: diversity, divergence, and introgression estimated using low-coverage genome sequencing. *PLoS Genetics*, **5**, e1000550.
Kulikova IV, Zhuravlev YN, McCracken KG, Haukos DA (2004) Asymmetric hybridization and sex-biased gene flow between Eastern Spot-billed Ducks (*Anas zonorhyncha*) and Mallards (*A. platyrhynchos*) in the Russian Far East. *The Auk*, **121**, 930–949.
Levin DA, Ortega JF (1996) Hybridization and the extinction of rare plant species. *Conservation Biology*, **10**, 10–16.
Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.
Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, **21**, 936–939.
Machado CA, Haselkorn TS, Noor MAF (2007) Evaluation of the genomic extent of effects of fixed inversion differences on intraspecific variation and interspecific gene flow in *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, **175**, 1289–1306.
Martin SH, Davey JW, Jiggins CD (2015) Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Molecular Biology and Evolution*, **32**, 244–257.
Martinsen GD, Whitham TG, Turek RJ, Keim P (2001) Hybrid populations selectively filter gene introgression between species. *Evolution*, **55**, 1325–1335.
Melo-Ferreira J, Boursot P, Suchentrunk F, Ferrand N, Alves PC (2005) Invasion from the cold past: extensive introgression of mountain hare (*Lepus timidus*) mitochondrial DNA into three other hare species in northern Iberia. *Molecular Ecology*, **14**, 2459–2464.
R Development Core Team (2008) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
Raj A, Stephens M, Pritchard JK (2014) fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, **197**, 573–589.
Renaut S, Grassa CJ, Yeaman S *et al.* (2013) Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications*, **4**, 1827.
Rhymer JM, Simberloff D (1996) Exinction by hybridization and introgression. *Annual Review of Ecology and Systematics*, **27**, 83–109.
Rieseberg LH (2001) Chromosomal rearrangements and speciation. *Trends in Ecology & Evolution*, **16**, 351–358.
Rieseberg LH, Soltis DE, Palmer JD (1988) A molecular reexamination of introgression between *Helianthus annuus* and *H. bolanderi* (Compositae). *Evolution*, **42**, 227–238.
Safford H, Viers J, Harrison SP (2005) Serpentine endemism in the California flora: a database of serpentine affinity. *Madroño*, **54**, 222–257.
Sambatti JBM, Strasburg JL, Ortiz-Barrientos D, Baack EJ, Rieseberg LH (2012) Reconciling extremely strong barriers with high levels of gene exchange in annual sunflowers. *Evolution*, **66**, 1459–1473.
Secondi J, Faivre B, Bensch S (2006) Spreading introgression in the wake of a moving contact zone. *Molecular Ecology*, **15**, 2463–2475.
Takayama K, Kajita T, Murata J, Tateishi Y (2006) Phylogeography and genetic structure of *Hibiscus tiliaceus*–speciation of a pantropical plant with sea-drifted seeds. *Molecular Ecology*, **15**, 2871–2881.
Van der Auwera GA, Carneiro MO, Hartl C *et al.* (2002) *From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline*. John Wiley & Sons Inc, Hoboken, New Jersey.
Vilà M, Weber E, Antonio CMD (2000) Conservation implications of invasion by plant hybridization. *Biological Invasions*, **2**, 207–217.
Wahlund S (1928) Zusammensetzung von Populationen und Korrelationserscheinungen vom Standpunkt der Vererbungslehre aus betrachtet. *Hereditas*, **10**, 65–106.
Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.
White M (1978) *Modes of Speciation*. W. H. Freeman and Company, San Francisco, California.
Wickham H (2009) *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York, New York.
Yatabe Y, Kane NC, Scotti-Saintagne C, Rieseberg LH (2007) Rampant gene exchange across a strong reproductive barrier between the annual sunflowers, *Helianthus annuus* and *H. petiolaris*. *Genetics*, **175**, 1883–1893.

G.L.O designed the experiment, collected the data, analyzed the data and wrote the paper. G.J.B. provided *H. annuus* sequence data and helped with analyses. L.H.R helped design the experiment and helped write the paper. All authors provided feedback for the final version of the manuscript.

Data accessibility

DNA sequences: NCBI SRA: SRP062491.

All raw sequence data are uploaded to the NCBI SRA (SRP062491). All custom scripts used in the analysis are included in Appendix S1 (Supporting information).

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 FastStructure plots for K = 1–10.

Fig. S2 ADMIXTURE plots for K = 1–10.

Fig. S3 Cross-validation scores for ADMIXTURE analyses.

Fig. S4 Patterson's D and f_d across the genome using all samples in 10 Mb windows.

Fig. S5 Comparing recombination rate with (a) Patterson's D, (b) f_d, (c) f_d for windows with positive D scores.

Fig. S6 Patterson's D scores for subsampled results.

Fig. S7 Patterson's D scores in California *Helianthus annuus* samples by latitude and collection date.

Fig. S8 Comparison of Patterson's D scores and latitude for *Helianthus bolanderi–exilis* samples.

Table S1 Sample information by population and by individual.

Table S2 Weir and Cockerham F_{ST} between all pairs of populations of *Helianthus bolanderi–exilis* and *Helianthus annuus*.

Table S3 Patterson's D and f_d for each chromosome and all results for subsampled ABBA–BABA tests.

Table S4 Results for all partitioned D tests.

Appendix S1 All custom scripts used in this manuscript, including a description of what each individual script does.

Appendix S2 Detailed explanation of all D partitioned results.