1-2021

Patterns of genetic variation in a prairie wildflower, Silphium integrifolium, suggest a non-prairie origin and locally adaptive variation

Andrew R. Raduski
University of Minnesota

Adam Herman
University of Minnesota

Cloe Pogoda
University of Colorado

Kevin M. Dorn
USDA-ARS

David L. Van Tassel
The Land Institute

See next page for additional authors

Raduski, Andrew R.; Herman, Adam; Pogoda, Cloe; Dorn, Kevin M.; Van Tassel, David L.; Kane, Nolan; and Brandvain, Yaniv, "Patterns of genetic variation in a prairie wildflower, Silphium integrifolium, suggest a non-prairie origin and locally adaptive variation" (2021). *Publications from USDA-ARS / UNL Faculty*. 2515.
https://digitalcommons.unl.edu/usdaarsfacpub/2515

American Journal of Botany
Botanical Society of America
leading scientists and educators since 1893

RESEARCH ARTICLE

Patterns of genetic variation in a prairie wildflower, *Silphium integrifolium*, suggest a non-prairie origin and locally adaptive variation

Andrew R. Raduski[1], Adam Herman[1], Cloe Pogoda[2], Kevin M. Dorn[3], David L. Van Tassel[4], Nolan Kane[2], and Yaniv Brandvain[1,5]

Manuscript received 25 June 2020; revision accepted 28 September 2020.

[1] Department of Plant and Microbial Biology, University of Minnesota Twin Cities, 140 Gortner Laboratory, 1479 Gortner Avenue, St. Paul, MN 55108, USA

[2] Ecology and Evolutionary Biology Department, University of Colorado, Ramaley N122, 1900 Pleasant Street, Boulder, CO 80309-0334, USA

[3] USDA-ARS, 1701 Centre Ave, Fort Collins, CO 80526, USA

[4] The Land Institute, 2440 E. Water Well Road, Salina, KS 67401, USA

[5] Author for correspondence (e-mail: ybrandva@umn.edu)

Citation: Raduski. A. R., A. Herman, C. Pogoda, K. M. Dorn, D. L. Van Tassel, N. Kane, and Y. Brandvain. 2021. Patterns of genetic variation in a prairie wildflower, Silphium integrifolium, suggest a non-prairie origin and locally adaptive variation. *American Journal of Botany* 108(1): 145–158.

doi:10.1002/ajb2.1603

PREMISE: Understanding the relationship between genetic structure and geography provides information about a species' history and can be used for breeding and conservation goals. The North American prairie is interesting because of its recent origin and subsequent fragmentation. *Silphium integrifolium*, an iconic perennial American prairie wildflower, is targeted for domestication, having undergone a few generations of improvement. We present the first application of population genetic data in this species to address the following goals: (1) improve breeding by characterizing genetic structure and (2) identify the species geographic origin and potential targets and drivers of selection during range expansion.

METHODS: We developed a reference transcriptome as a genotyping reference for samples from throughout the species range. Population genetic analyses were used to describe patterns of genetic variation, and demographic modeling was used to characterize potential processes that shaped variation. Outlier scans for selection and associations with environmental variables were used to identify loci linked to putative targets and drivers of selection.

RESULTS: Genetic variation partitioned samples into three geographic clusters. Patterns of variation and demographic modeling suggest that the species origin is in the American Southeast. Breeding program accessions are from the region with lowest observed genetic variation.

CONCLUSIONS: This prairie species did not originate within the prairie. Breeding may be improved by including accessions from outside of the germplasm founding region. The geographic structuring and the identified targets and drivers of adaptation can guide collecting efforts toward populations with beneficial agronomic traits.

KEY WORDS Asteraceae; de novo domestication; population genetics; prairie; Silphium; transcriptome.

Understanding the forces that influence genetic structuring of species is a fundamental goal of population genetics. It is not yet well understood how potentially drastic recent anthropogenic habitat changes can affect already established patterns of genetic structuring, especially in long-lived perennial plants. The North American prairie was once the largest vegetative region on the continent. It has served as a cultural identity reference point and the subject of many grassland ecological studies (Weaver, 1954; Lauenroth, 1979; Matthews, 1988; Wilson and Hartnett, 1998; Samson et al., 2004; Knapp et al., 2008). The overwhelming majority of the prairie has since been altered by human disturbance with some estimates of range decline since European settlement as high as 99.9% (Samson and Knopf, 1994). Conservation efforts of this endangered habitat have focused on protection of virgin prairie fragments and restoration of altered landscapes by human-directed reseeding of prairie plants. The restoration approach is far more common, in part, because of the scarcity of available virgin prairie. Having knowledge about how genetic variation is distributed across natural landscapes can be helpful toward deciding which virgin prairies to prioritize or which geographic ecotypes of plant species to place in prairie restorations.

Surprisingly little is known about the evolutionary history of common North American prairie forbs from population genetic data, and it is not currently known whether there are general patterns across species of directional range expansion following glaciation or locations of geographic sources of ancestral variation. Changes in species composition have drastically altered the biome as C_4 grasses became dominant on the landscape over the past few million years (Fox and Koch, 2003, 2004). Data from fossil pollen deposits have suggested that now common prairie wildflower families, such as the Asteraceae, were likely not plentiful on the prairie landscape at the time that C_4 grasses became dominant (Williams et al., 2004). Understanding the evolutionary and demographic history of how a species has come to inhabit its range can allow us to better understand the present relationship between a species's genetic variation and geography. Currently, the extent to which recent habitat fragmentation affects genetic signatures of past range expansion and how a species life history affects the pattern is not well understood. Perennial plants may be buffered more to a degree than annuals from genetic diversity losses brought on by habitat fragmentation.

Prairie perennial plants have become the inspirations and models for neo-domestication efforts to develop sustainable agriculture. Roots of perennial plants in the soil protect against soil erosion, promote a diverse soil microbiome, and act as a carbon sequestration resource. Species of the genus *Silphium* (Asteraceae) are deep-rooted, long-lived, perennial wildflowers with flowers pollinated by bees and seeds dispersed by wind that are found throughout the United States, east of the Rocky Mountains. Common *Silphium* species are best known for being charismatic flora of the North American prairie and their ability to tolerate extreme abiotic stresses (Weaver et al., 1935; Leopold, 1968). Recently, there have been efforts to better incorporate the ecological benefits of *Silphium* species into improving habitat restoration and agricultural practices (Van Tassel et al., 2017). Because of these benefits, combined with high biomass accumulation requiring relatively few inputs, species of *Silphium* show promise for forage and biogas production (Ustak and Munoz, 2018; Wever et al., 2019) or as a perennial oil seed crop (Van Tassel et al., 2017). Additionally, the incorporation of native wildflowers in prairie restoration projects or as buffer strips among agriculture fields are thought to improve pollinator abundance (Isaacs et al., 2009). *Silphium* species are known to attract beneficial insects (Fiedler and Landis, 2007) and are currently being evaluated for such conservation practices because they are large plants, competitive with weeds after established, have potential for economic return (seed or bioenergy) and breeding programs may improve seedling vigor compared to other wildflowers (Butters et al., 2019, unpublished poster, Entomology 2019 conference, St. Louis, MO, USA). Despite increased interests across applied settings, relatively little is known about patterns in the distribution of genetic variation in long-lived perennial prairie genera such as *Silphium*, or how recent habitat fragmentation and degradation have affected levels of variation in natural populations.

As there has been a recent movement to domesticate *S. integrifolium* and *S. perfoliatum*, understanding the underlying extent and distribution of genomic variation within and among populations will help us better understand how recent habitat fragmentation has affected the genetic signature of range expansion and also support ongoing applied *Silphium* research. Breeding trials to develop *S. perfoliatum* as a potential cover crop began in the 1950s and continue today with additional effort to develop it as a biogas resource (Stanford, 1990; Wever et al., 2019). Most recently, *S. integrifolium* has been targeted for development as an oilseed crop, in part, because the fatty acid composition profile in its seeds is similar to that in sunflower seeds (Reinert et al., 2019). Accessions used in *S. perfoliatum* breeding have little genetic diversity from (often) unknown geographic origins (Wever et al., 2019). To date, most *S. integrifolium* breeding progress has been made from accessions that originated within a small geographic area (Kansas, USA) (Vilela et al., 2018), with a general lack of genomic resources (Van Tassel et al., 2017). The extent to which breeding populations are representative of species-wide diversity has not been genetically evaluated. However, if populations that produced *S. integrifolium* breeding accessions are not representative of species-wide diversity, greater improvements may be achieved by incorporating accessions with unique diversity similar to the practice of incorporating new, useful genetic variation from wild relatives into crop breeding programs (McCouch et al., 2013). Additionally, the choice of locations for field trials may be informed by a better understanding of the relationship between natural genetic variation and geography.

Presently, *Silphium* genomic resources are lacking, in part, because of the difficulty of assembling its large, repetitive genome. All *Silphium* species are diploid ($2n$ = 14) with genome sizes that are nearly triple that of their close relative sunflower (*Helianthus annuus*) (Bai et al., 2012). We utilized transcriptome resequencing of *S. integrifolium* to characterize and quantify functional genetic variation throughout the species range. We assembled a reference transcriptome of a diverse set of tissues from a single individual and use the reference to identify variants from a panel of resequenced young leaf transcriptomes.

Here, we characterized genetic variation among samples of *S. integrifolium*. We used this variation to investigate the distribution of neutral genetic variation across the species range while maintaining an interest in identifying a possible region of geographic origin of the species. In addition to uncovering genome-wide population structure, we hypothesized that populations of this widespread species are locally adapted to their unique environments, and thus, quantifying the relationship between genetic variation and geography is important. Therefore, we next sought to identify genes whose patterns of genetic variation have been shaped by natural selection. Such environmentally associated variants highlight traits experiencing geographically variable selection in this species and could identify important variation to include in the germplasm for the breeding program.

MATERIALS AND METHODS

De novo transcriptome sampling, sequencing, and assembly

Five tissues were collected from a single *S. integrifolium* plant from the breeding program of The Land Institute in Salina, Kansas, United States. Root, crown, flower, leaf, and stem tissues were collected and immediately frozen in liquid nitrogen. Flash-frozen samples were transported to the University of Minnesota on dry ice where total RNA was extracted using Qiagen RNeasy kits (Germantown, Maryland USA). Pooled RNA-sequencing (RNA-seq) stranded libraries were constructed by the University of Minnesota Genome Center and sequenced on two lanes of an Illumina (San Diego, California USA) HiSeq instrument (2 × 125 paired-end reads per lane).

Raw sequence reads were trimmed of low-quality sites and adaptor sequences using BBDuk (Bushnell, 2018) and de novo assembled into a reference transcriptome with Trinity version 2.1.1 (Haas et al., 2013) and non-default settings: —SS_lib_type_type RF —min_contig_length 300 —min_kmer_cov 2 —min_glue 5.

Assembled transcriptome completeness and annotation

The completeness of the initial Trinity *S. integrifolium* assembly was measured by assessing the number of benchmark universal single-copy orthologs with the program BUSCO v3.0.2 with the *embryophyta_odb9* gene set (Simão et al., 2015). The longest isoform of each transcript was used for BUSCO searches.

The number of nearly full-length protein-coding transcripts was counted by querying the Swiss UniProt protein database (Boutet et al., 2016) with the full set of transcripts using BLASTx with an e-value cutoff of 1e-20 (Madden, 2013). For each unique result from the protein database, the percentage of matching sequence length of the best matching transcript was recorded. Transcripts that matched a database entry over more than 80% of its length were classified as nearly full-length transcripts.

Transcript abundance was estimated with the RSEM software (Li and Dewey, 2011) after mapping cleaned sequence reads to the Trinity assembly using Bowtie v1.1.2 (Langmead and Salzberg, 2012). Gene abundance matrices were built after calculating the number of gene transcripts per million sampled transcripts (TPM) using Trinity scripts. Transcript abundance matrices per gene were built to determine the isoform prevalence within each gene. The median contig length (N50) of the transcriptome assembly was calculated using all transcripts (E100N50) and only transcripts with expression levels above the 90th percentile (E90N50). Although the N50 statistic is widely reported without regard for expression level, it may be an inaccurate representation of transcriptome assembly because of a bias from low-expressed transcripts.

The longest isoform of each contig was functionally annotated using the standard Trinotate pipeline (Bryant et al., 2017). Possible protein-coding sequences were identified using TransDecoder (Haas and Papanicolaou, 2017). Putative proteins longer than 100 amino acid residues were kept. BLASTx searches using transcripts and BLASTp searches using predicted amino acids as queries were performed against the Uniprot protein sequence database (Madden, 2013; Boutet et al., 2016).

Reference transcriptome filtering

The Trinity assembly was further filtered to create a subset of contigs to serve as a genotyping reference for downstream population genetic analyses. We initially retained only contigs that had a significant BLAST result for a plant after using the longest isoform of each contig as a query to the Swiss UniProt database and the NCBI non-redundant nucleotide database with BLASTp and BLASTn searches, respectively. E-value cutoffs of 1e-20 (BLASTp) or 1e-30 (BLASTn) were used. Plant contigs were identified by keeping queries with a best BLAST result belonging to a genus on the NCBI viridiplantae taxonomy list (Sayers et al., 2009). The completeness of the filtered contig set was assessed by again performing a BUSCO analysis of TransDecoder predicted amino acids with the *embryophyta_odb9* gene set. The plant-filtered contig set was further reduced to include only contigs that were inferred to be single-copy orthologous genes in both *S. integrifolium* and *H. annuus* with OrthoFinder (Emms and Kelly, 2015). The resulting set of contigs served as a genotyping reference for downstream resequencing of samples.

Transcriptome resequencing

Fresh young leaf tissue was collected and flash frozen from 68 wild *S. integrifolium* plants from 39 sampling locations and one *S. laciniatum* plant to be used as an outgroup (Appendix S1). We often have a single or few individuals from each sampling location; however, sampling locations are often clustered near each other. Self-incompatible species such as *S. integrifolium* are expected to have less genetically structured populations and, thus, performing genetic structure analyses with few individuals from many sites across a species range may result in an underestimate of fine-scale structure (Charlesworth, 2003). All plants were grown from wild-collected seed in a common garden at The Land Institute in Salina, Kansas, United States. Pooled RNAseq stranded libraries were constructed by the University of Minnesota Genome Center and sequenced using an Illumina (San Diego, CA, USA) HiSeq instrument (2 × 125 PE reads per lane).

Variant-calling workflow

Reads from each resequenced leaf transcriptome were trimmed of adaptor sequences using scythe v0.991 (Buffalo, 2014) and aligned to the genotyping reference using STAR v2.5.2b (Dobin et al., 2013). Reads were aligned in a two-pass mode to accommodate the possible mapping of isoforms across splice junctions. PCR duplicates were removed from each alignment using Picard tools (Broad Institute, 2018). Sample genotypes at variant single nucleotide polymorphisms (SNPs) and nonvariant sites were called using Freebayes with all samples assigned to a single population (Garrison and Marth, 2012 [Preprint]). A minimum site depth of 10 reads was required within each sample for a site to be included in genotype calling.

Population structure and genetic diversity analyses

A principal component analysis (PCA) was performed on *S. integrifolium* genotypes to assess genetic population structure across the species range. Single nucleotide polymorphisms (SNPs) were filtered to remove rare variants (minor allele frequency [MAF] ≤ 0.05) and sites in linkage disequilibrium (LD) (LD threshold = 0.2, missingness = 0%) before the PCA. We performed a Procrustes transformation of sample principal components (PC1 and PC2) coordinates onto their sampling latitude and longitude coordinates to better visualize the relationship between genetic variation structuring and geographic space (Wang et al., 2010).

The relative effect of isolation by distance on the observed population structure patterns was estimated using conStruct (Bradburd et al., 2018). Similar to other STRUCTURE-like programs (Pritchard et al., 2000), conStruct assigns a proportion of ancestry from different genetic groups (K) to individuals. Additionally, conStruct can incorporate a spatial element into its model to account for isolation by distance. We ran spatial and nonspatial models over a range of underlying genetic groups (K = 1 through K = 6), and found spatial models always had higher predictive accuracy. We ran multiple spatial models (K = 1 through K = 6) to identify the best K value with a scree-plot of predictive accuracy. For each value of K, we ran four independent runs with a 1000 iteration chain length. Five independent runs, each with four MCMC chains of 5000 iterations were then performed with the inferred optimal *K* value.

Genetic relationships among *S. integrifolium* samples were inferred using SVDQuartets (Chifman and Kubatko, 2014). SVDQuartets estimates an unrooted phylogeny describing the relationships between sets of four taxa. The correct phylogeny topology under a coalescent process is inferred by singular value decomposition of the genetic distance matrix containing the four taxa. Multiple quartet phylogenies are then assembled into a single phylogeny that contains all taxa. A matrix of biallelic SNPs was used with exhaustive quartet sampling, 100 bootstrap replicates, and the Quartet MaxCut tree assembly method to produce an unrooted phylogeny within PAUP* version 4.0a (build 166; Swofford, 2003). A single individual from Illinois was removed from the SNP data matrix before running SVDQuartets because we observed evidence that it was of unusually high intraspecific admixed ancestry.

Results from PCA, conStruct, and SVDQuartets indicated that the major axes of genetic variation separate *S. integrifolium* samples into three clusters that broadly correspond to geographic regions. Hereafter, we refer to these broad groupings as the eastern (Mississippi, Arkansas, Missouri, Wisconsin, and Illinois plants), southern (Texas and Oklahoma plants), and western (Kansas and Nebraska plants) regions. When describing finer-scale comparisons within regions, we identify groups of plants by their American state of sampling origin. These groups were not a priori assumed to be populations in the traditional population genetic sense.

The level of support for region-level relationships among groups within *S. integrifolium* across markers was assessed using TWISST (Martin and Van Belleghem, 2017). TWISST prunes a given phylogeny down to one tip from each user-defined taxon group and reports the topology of the pruned tree. The process is repeated, and support for each topology is recorded as the number of observations. A maximum likelihood phylogeny was inferred for each contig using RAxML v8.0 (Stamatakis, 2014), and each tip was assigned to a regional taxon inferred from the above clustering methods. The pruned topology most often observed for each contig's reconstructed phylogeny was recorded.

Genetic differentiation among sets of *S. integrifolium* plants grouped by their sampling state of origin was measured by calculating Wright's F_{ST} (Weir and Cockerham, 1984) between all pairwise combinations using a stringently filtered marker data set with the R package StAMPP (Pembleton et al., 2013). Retained SNP markers were biallelic in *S. integrifolium*, had a Phred quality score ≥20, had an allele balance within heterozygotes between 0.25 and 0.75 or within homozygotes less than 0.05, had average sequencing depth greater than 10 and less than 200, and MAF ≥0.05. One hundred bootstrap replicates were performed across loci to calculate 95% confidence intervals around F_{ST} estimates and to check whether our grouping of plants by sampling state of origin at least reasonably approximates biological populations.

Average numbers of nucleotide differences per base pair at 4-fold degenerate sites were calculated between all samples to measure genetic diversity (Nei, 1987). We organized the diversity measurements into within (π) and between (D_{xy}) sets of plants grouped by their sampling state of origin. Variation was calculated across contig bootstrap replicates to calculate 95% confidence intervals around mean values. Tajima's *D* was calculated within each group of plants on every contig using vcftools (Danecek et al., 2011). The level of partitioning of genetic variance was measured using an AMOVA performed on genetic distances (Nei, 1972) between individuals with states and regions as hierarchical levels of stratification with the R packages StAMPP and pegas (Pembleton et al., 2013; Paradis, 2010). Statistical significance of variance partitioning among stratification levels was assessed with 10,000 random permutations.

We found that samples from Illinois exhibited patterns of nucleotide variation, Tajima's *D*, and inferred ancestry that were unexpected within the larger contextual patterns of each metric across the species landscape. We performed the three population f_3 admixture tests in TreeMix to test the hypothesis that the Illinois population (IL) is the result of past admixture between Wisconsin (WI) and a third population (*X*) instead of the null expectation that the relationships among the sampled populations are the product of a tree-like branching process (Reich et al., 2009; Pickrell and Pritchard, 2012). Each test took on the form {f_3 IL; WI, *X*}, where evidence that IL is the result of admixture between WI and *X* is given by a significantly negative test statistic. For all identities of *X*, *Z*-scores and standard errors of estimates were reported from jackknifing 50 SNP blocks. We subsequently removed all Illinois samples from our demographic inference modeling so as to not confuse historical processes with potential anthropogenic movement.

Data from the three-dimensional joint site frequency spectra were used to infer general branching patterns of *S. integrifolium* regional taxa using Moments (Jouganous et al., 2017). For three-dimensional spectra, the tested branching scenarios were those that would result in the tree topologies (East, (South, West)), (South, (East, West)), and (West, (South, East)). For each branching pattern scenario, we fit demographic models with no migration, symmetric migration between each taxa pairs, and asymmetric migration rates. The fit of each model was assessed across 100 replicates using different starting parameters selected from bounded limits decided by trial and error. The likelihood of each model was optimized using Powell's method, and the quality of each model's fit to the data was assessed with Akaike information criterion (AIC) scores. After the best-fitting model was selected for each scenario, uncertainty estimates around each parameter were calculated using the Godambe information matrix with 100 bootstrapped spectra (Jouganous et al., 2017).

Scans for selection

To identify loci possibly under divergent selection, we used BayeScan (Foll and Gaggiotti, 2008; Foll, 2012) to find statistical outlier SNPs that significantly deviate from neutral expectations. We calculated global F_{ST} scans across sites between regions using the same marker set used to perform an AMOVA. Default BayeScan model parameters were used (burn-in length = 50,000, thinning interval = 10, sample size = 5000, number of pilot runs = 20, pilot run length = 5000). Annotation of resulting outlier loci was done using BLASTx, BLASTp, and eggNOG searches using Trinotate scripts. We then sought to identify environmental variables that may be associated with the selective pressures acting on each outlier protein. We identified environmental variables that had significant associations with genetic markers within each protein after accounting for the coancestry of samples using BayEnv2 (Günther and Coop, 2013). Environmental variables from the geographic midpoint of each population were extracted from the ClimateNA data set (Wang et al., 2016). Brief descriptions of the climatic variables used are provided in Appendix S2. For each BayeScan identified protein, the Bayes factor, absolute Spearman's rank, and absolute Pearson correlation coefficients with the largest quantile were used to identify putative environmental variables driving local adaptation.

RESULTS

De novo transcriptome and assembly

Trinity assembled 541,120,911 paired-end sequencing reads into 280,573 transcripts and 166,458 contigs, totaling over 246.9 Mb of assembled nucleotides. The median contig length of all transcripts (E100N50) was 1159 bp. The median contig length of transcripts within the top 90th percentile of expression (E90N50) was 2014 bp. The filtered assembly with the longest isoform of each contig totaled 123.5 megabases. A BUSCO analysis of the longest isoform of each assembled contig found that 80.7% of the 1440 genes in the *embryophyta_odb9* data set were found in complete form (71.3% as single copy and 9.4% as duplicated). BLASTx searches of the Swiss UniProt database identified 13,143 unique protein results. Of those, 8418 (64.0%) were found as full-length or nearly full-length transcripts in the full Trinity transcriptome assembly.

Reference transcriptome filtering

After filtering the full Trinity assembly for contigs with a BLASTp or BLASTn result from genera on the NCBI viridiplantae taxonomy list, 51,238 contigs remained. Results of a BUSCO analysis of this filtered contig set were similar to the full assembly analyses (80.7% of 1440 BUSCO genes found complete, of which, 71.4% were single copy and 9.3% were duplicated). The plant-filtered contig set was further reduced to 10,575 contigs that contained orthologous genes that were inferred to be single copy in *Silphium* and *H. annuus*. This final contig set was used as a reference to align resequencing reads and to call variants.

Transcriptome resequencing and variant calling

Leaf transcriptome resequencing data mapped to 7718 of the 10,575 contigs in the genotyping reference set. Genotyping was restricted to contigs with mapped reads and covered 12,410,795 total sites (852,148 SNPs) across all samples. One *S. integrifolium* sample from Missouri was removed from downstream analyses because of low average sequencing depth (<10) and a high rate of missing markers (>60%).

Population structure and genetic diversity analysis

PC1 and PC2 explained 13.79% of the variation in our *S. integrifolium* samples. Genotypes formed three distinct groups on the first two PC axes (Fig. 1B). The groups broadly consisted of plants from the eastern (Arkansas, Illinois, Mississippi, Missouri, and Wisconsin), southern (Oklahoma and Texas), and western (Kansas and Nebraska) regions of the species range. Eastern and western regions were separated along the first PC axis. The southern region was separated from eastern and western regions along the second PC axis.

Our conStruct analyses that incorporated isolation by distance into population structure (spatial models) always performed better than nonspatially aware models. Across 10-fold cross-validation runs for each value of *K*, the improvement of spatial over nonspatial models ranged from 291.79 to 54.52 log-likelihood units. Results from spatially aware runs with different numbers of *K* indicated that relatively little predictive power was gained after increasing the number of genetic layers past *K* = 3. The patterns of assigned ancestry across all samples supported the PCA pattern of subdivision. Within the eastern region, samples overwhelmingly consisted of ancestry from a single dominant genetic layer. Samples from the southern and western regions are inferred to be best described as a mixture of ancestrty, with both regions being made

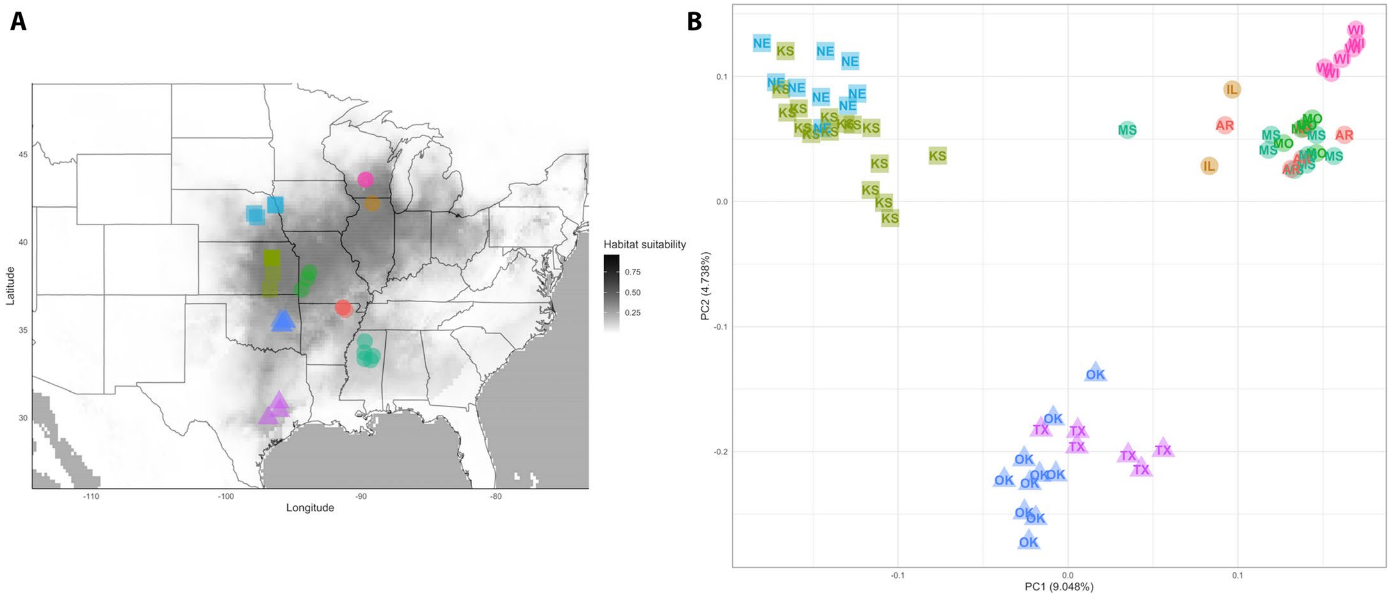

FIGURE 1. (A) Suitable *S. integrifolium* range estimate and current study sampling locations. Range estimate is shaded by MaxEnt estimated habitat suitability (see Appendix S8 for method details). Sampling localities are colored according to their American state. (B) Results from a principal components analysis of *S. integrifolium* genotypes. The two major axes of variation distinguish samples into three clusters that correspond to regional (eastern, southern, and western) geography. Coloring on principal component biplot corresponds to same states in panel A.

up of a different dominant ancestry layer (Fig. 2A, B). In western samples, the second largest ancestry layer is the dominant layer in southern samples. The reciprocal is true in most southern samples.

Reconstruction of the phylogenetic relationships of samples using SVDQuartets resulted in strong geographic clustering with 100% bootstrap support for all nodes in the phylogeny (Fig. 2C). All but one sample from the eastern region formed a large clade, with most samples from within a population forming smaller clades. Mississippi samples exhibited a polyphyletic pattern throughout the larger eastern clade, with various samples clustered with Missouri or Arkansas samples. A single Mississippi sample was reconstructed to be outside of a larger clade containing all other *S. integrifolium* samples. Southern samples are paraphyletic with most Oklahoma samples forming a clade that is sister to a single Texas sample. Most Texas samples formed a clade that is sister to a larger clade that contained a single Oklahoma sample and a larger clade of all western samples (Fig. 2).

To investigate the level of support for the general SVDQuartets phylogeny across markers, we performed a TWISST analysis. Each maximum likelihood phylogeny was rooted with *S. laciniatum* as the outgroup, and the region-level topology of *S. integrifolium* taxa phylogeny was recorded. For each taxa in a tree, the identity of its sister tip was recorded. If the nearest sister was a node, a descendant of the sister node was randomly sampled and given a weight proportionate to the number of descendants of the sister node. Confidence around these measurements was estimated with 25,000 bootstraps across contig phylogeny topologies. Similar to the SVDQuartets phylogeny topology, southern and western regions were sister to each more often than either was sister to the eastern region. The eastern region was found to be sister to western and southern regions an approximately equal number of times across markers (Appendix S3). For each focal taxon's set of sister taxa, a χ^2 test was performed assuming a star phylogeny, i.e., equal probabilities of observing each sister taxon. Equal probabilities of observing each sister taxon was rejected for the western and southern region focal taxa ($p < 0.05$ Bonferroni-corrected for multiple comparisons).

Across all comparisons, pairwise F_{ST} ranged from 0.0340 to 0.2204. Differentiation patterns between regions supported the structure inferred from PCA. As a general pattern, differentiation within regions was approximately $F_{ST} \leq 0.10$ regardless of geographic distance separating samples (Fig. 3). The rates of decay of the covariance between allele frequency and geographic distance within each genetic layer (conStruct's αD parameter, $K = 3$)

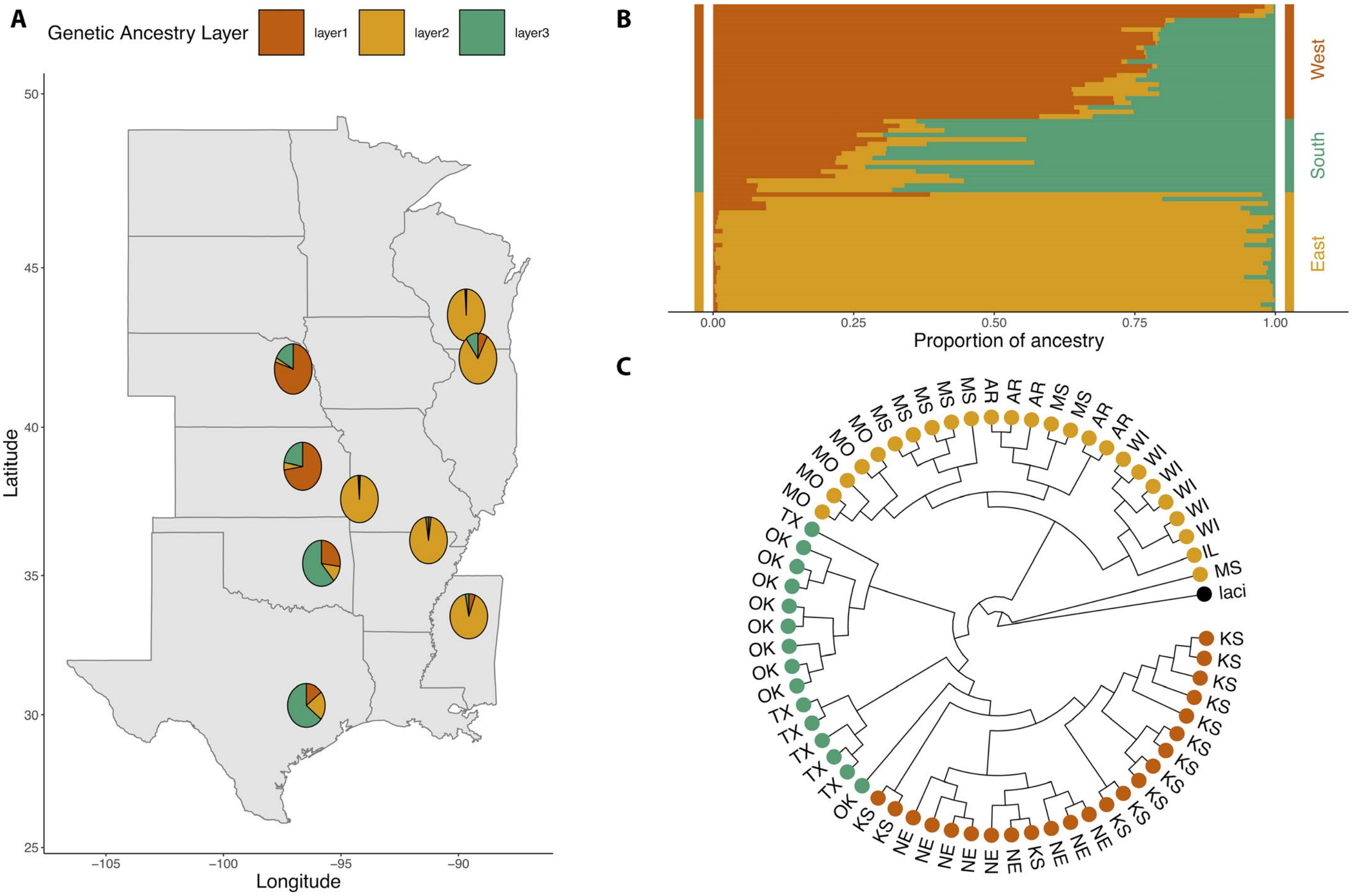

FIGURE 2. *Silphium integrifolium* samples comprise three genetic groups. (A, B) Inferred ancestry results of *Silphium integrifolium* samples from a conStruct model that includes spatial information and *K* = 3 (Bradburd et al., 2018). (A) Average amounts of inferred ancestry from three genetic layers for each population are displayed as pie charts. (B) Proportion of inferred ancestry of each genetic layer for all sampled plants. (C) SVDQuartets reconstruction of phylogenetic relationships between *S. integrifolium* samples, rooted on single *S. laciniatum* sample. Bootstrap node support values are not shown; every node has 100% bootstrap support.

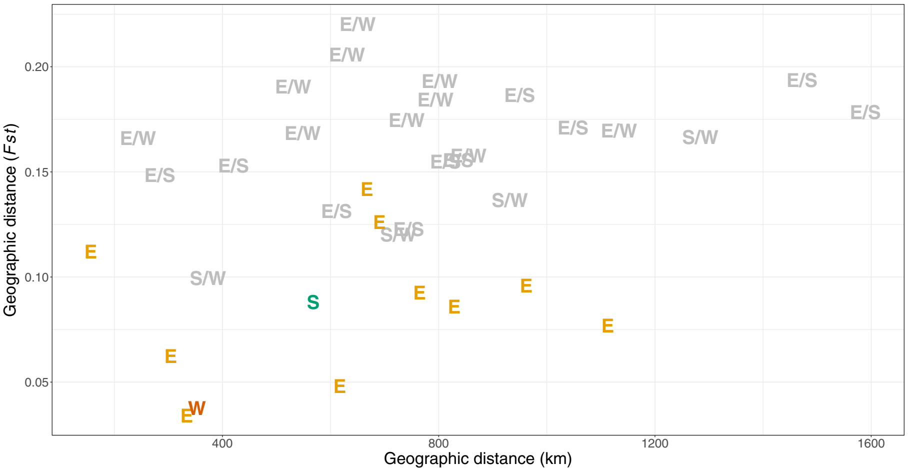

FIGURE 3. Relationships between geographic distance and genetic differentiation. Geographic distance was measured as Euclidean distance between the sample midpoints of sets of samples grouped by sampling state of origin. Differences in color reference within and between region comparisons; colored letters (E, S, W) represent within region comparisons and gray letter pairs (E/S, E/W, S/W) represent between region comparisons. There is no significant effect of geographic distance on genetic differentiation within each type of comparison (within and between regions, $p > 0.05$).

ranged from 0.001 to 0.005 suggesting little genetic differentiation above what is expected under isolation by distance within genetic layers. Differentiation between regions was generally higher than within-region differentiation with increased differentiation observed between geographically distant samples (Fig. 3); however, the relationship between geographic distance and differentiation was not significant within each type of comparison (within region and between regions). The three highest within-region comparisons all involved the suspect IL locality, likely reflecting the recent admixture in this sample. The highest levels of differentiation were observed between eastern and western regions and between eastern and southern regions. The greatest divergence in allelic frequency was between Nebraska (western region) and Illinois (eastern region) populations ($F_{ST} = 0.2204$). Our AMOVA results showed that genetic variation is strongly partitioned both among regions and among states of sampling within regions (Table 1).

The average numbers of pairwise nucleotide differences per base pair at 4-fold degenerate sites (π) between samples within regions were calculated to estimate genetic diversity (Fig. 4B; Appendix S4). The eastern region had the highest mean genetic diversity of any group ($\pi = 0.01013$, 95% CI 0.00986–0.01041). The southern region had slightly lower genetic diversity ($\pi = 0.00982$, 95% CI 0.00952–0.01013). The western region had lowest genetic diversity ($\pi = 0.00894$, 95% CI 0.00865–0.00923).

Within the eastern and southern regions, samples from states with lower latitudes had higher π, consistent with those regions possibly serving as founder populations for northward expansion after glaciation (Fig. 4A; Appendix S4). Mississippi and Arkansas contained more genetic diversity than other eastern region state groupings ($\pi = 0.01018$, 95% CI 0.00987–0.01051 and $\pi = 0.00972$, 95% CI 0.00935–0.01008, respectively). Within the southern region, genetic diversity in Texas was greater than Oklahoma ($\pi = 0.0098$, 95% CI 0.00944–0.01017 and $\pi = 0.00949$, 95% CI 0.00918–0.0098, respectively). The reduction in genetic diversity found in Kansas and Nebraska ($\pi = 0.00846$, 95% CI 0.00815–0.00878 and $\pi = 0.00948$, 95% 0.00918–0.00978, respectively) is suggestive of a past bottleneck of *S. integrifolium* into the western region.

Patterns of nucleotide divergence between populations (D_{xy}) support the regional structure grouping of plants found through PCA and conStruct. The largest D_{xy} observed, on a regional scale, was between eastern and southern regions ($D_{xy} = 0.0105$, 95% CI 0.01021–0.0108), only slightly exceeding π in the diverse eastern region. The largest levels of divergence were found between Mississippi and Texas ($D_{xy} = 0.01069$, 95% CI 0.01036–0.01102), followed by Texas and Wisconsin ($D_{xy} = 0.01068$, 95% CI 0.01035–0.01102), and Mississippi and Nebraska ($D_{xy} = 0.01067$, 95% CI 0.01036–0.01098). Plants from the western region had the smallest divergence from southern region plants, suggesting that the western samples are a result of a founder event originating from the southern region (Figs. 3, 4C; Appendix S4).

Tajima's *D*, calculated within each state across contigs, was consistently negative within samples from eight of nine states (Fig. 5), suggesting population growth across the range. We note that inferred population growth runs counter to the known recent contraction of prairie species due to habitat loss and likely reflects population growth since the glacial retreat. A recent selective sweep can also cause negative Tajima's *D*; however, the genomic signature will be confined to genetic markers linked to the target of the sweep. Yet another alternative is this negative Tajima's *D* reflects population substructure within states; however, the distribution of pairwise π within states argues against this explanation. The genome-wide pattern of negative Tajima's *D* we observe is thus interpreted to be due to population expansion, a process that will affect markers across the genome. The western region, (Kansas and Nebraska), Oklahoma, and Mississippi samples had the most negative Tajima's *D* values. Illinois was the only grouping that had an average Tajima's *D* value greater than zero ($D = 0.201$), likely reflecting the excess of intermediate frequency variants induced by rampant recent admixture.

The relatively high genetic variation (π), elevated Tajima's *D* estimates, and amounts of mixed ancestry inferred by construct within the IL population led us to test for a history of admixture using the f_3 test for treeness (Reich et al., 2009). We assumed that the IL and WI samples have recently diverged from each other via a common ancestral population and thus tested the null hypothesis that a tree-like branching pattern would best describe the process that formed any trio of populations containing IL and WI in the form f_3\{IL; WI, X\}. We found significantly negative *Z*-scores for the f_3 test when the third population in the trio was any western or southern grouping (KS, NE, OK, or TX), but not any other eastern grouping (AR, MO, or MS) (Table 2). Overall, this pattern suggests that IL is the product of admixture between the general eastern region lineage and the western or southern regions lineages or the common ancestor of the two.

Data from the joint site frequency spectra was used to model the demographic history of *S. integrifolium* under three general scenarios. The scenarios differed from each other by the identities

TABLE 1. AMOVA based on hierarchical grouping stratification among regions, among populations within regions, and within populations. ***p* ≤ 0.01, ****p* ≤ 0.0001.

Sources of variation	df	SS	Variance component	Percent of variance
Regions	2	0.44905	0.00796**	29.4869
Populations	6	0.23872	0.00375***	13.9031
Within populations	58	0.88744	0.01530	56.610
Total	66	1.57522	0.027028	100.00

of the two regional taxa assumed to be sister to each other, e.g., (East, (West, South)), (South, (East, West)), and (West, (East, West)). Under each model, an ancestral population was forced to split into two lineages, with one lineage later experiencing a second split to produce three total taxa. The proportions of a lineage that went into a split at a branching event, timing of branching events, migration rates between all lineages, and population sizes were estimated for each model. Migration rates differed in each model to either be nonexistent, symmetric, or asymmetric.

Under all branching scenarios considered, models that included migration were preferred over no-migration models. Models with migration rates constrained to be symmetric were preferred over asymmetric migration models. The three different branching scenarios considered for three-dimensional spectra each resulted in similar maximum likelihoods and AIC scores (Fig. 6, Table 3). Overall, the best-fitting model produced the (East, (South, West)) topology with substantial population expansion in the South and West with migration throughout all populations. This model estimates that approximately 40% of the ancestral population split ≈ 3.7N_e generations ago to form the common ancestor of South and West regions. This lineage then approximately doubled in size before splitting again ≈ 1.6N_e generations ago to create the South and West regions. At the most recent branching event, the majority of the population (83.7%) split to form the South region, which has since decreased by nearly half. The West region's population size has increased nearly 4-fold since the most recent split. The population of the East region has been stable in size since its initial split from the common ancetor. Estimated migration rates are highest between East–South and South–West regions; however, confidence intervals around migration rates are wide. Parameter estimates under all models are reported in Table 3.

Scans for selection

We performed a transcriptome-wide scan for divergence outliers on a set of 41,159 high quality markers with BayeScan. We detected 47 SNPs that were F_{ST} outliers while using a maximum false discovery rate of 0.001. The outlier SNPs were spread across 40 contigs, 34 of which we were able to gather annotation information from BLAST and eggNOG database searches. The summarized annotation information for selected proteins is given in Appendices S5, S6, and S7. For each protein identified by BayeScan as containing an outlier SNP, we found the environmental variable that had the highest correlation with any SNP based on Bayes factor, Spearman rank correlations, or Pearson correlation coefficient quantiles using

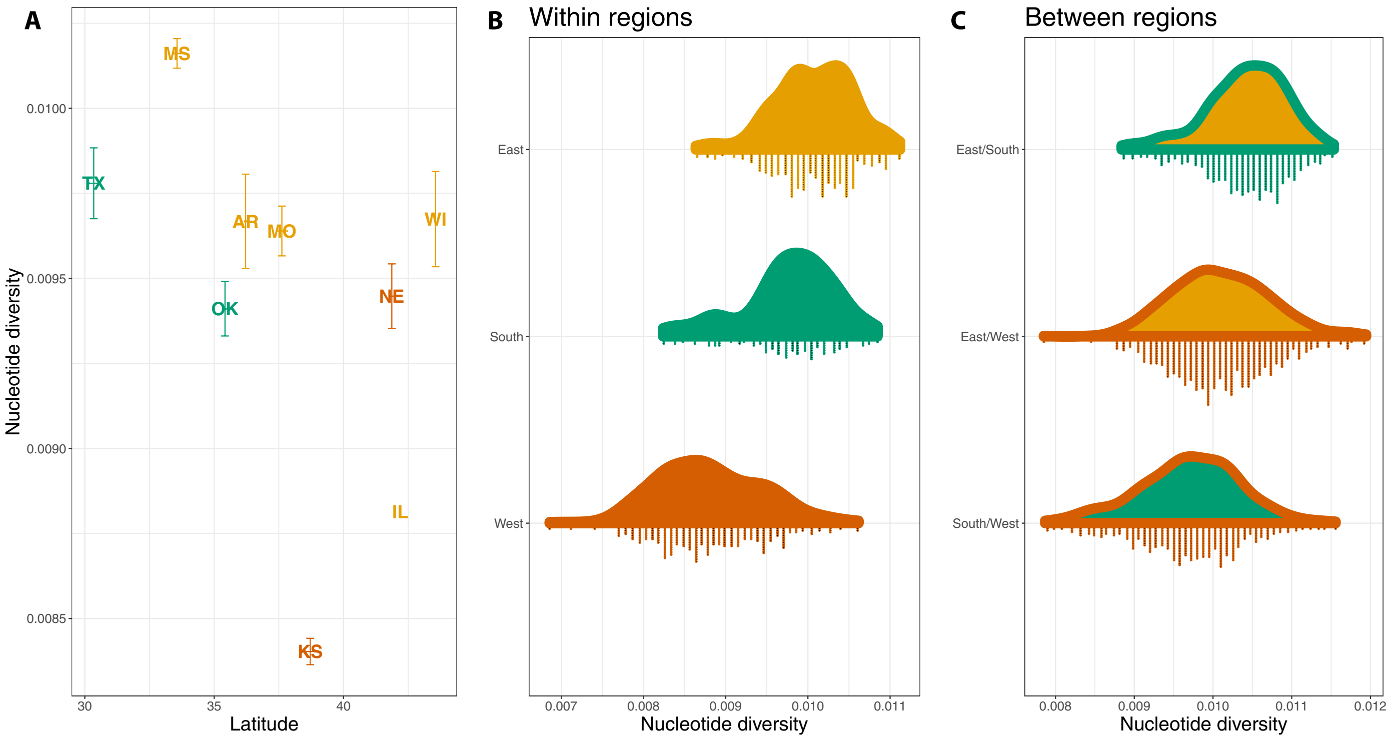

FIGURE 4. Distribution of nucleotide diversity at 4-fold degenerate sites. (A) Mean nucleotide diversity between samples grouped by sampling state of origin. For each state, latitude is calculated as the latitude midpoint of sampled plants. As a general pattern, diversity decreases with latitude. (B) Pairwise nucleotide diversity raincloud plots among samples within regions (π) and (C) among samples between geographic regions (D_{xy}). Each raincloud plot displays the nucleotide diversity density function above the observed raw data points. The outline and fill colors of the density distributions correspond to the regions being compared.

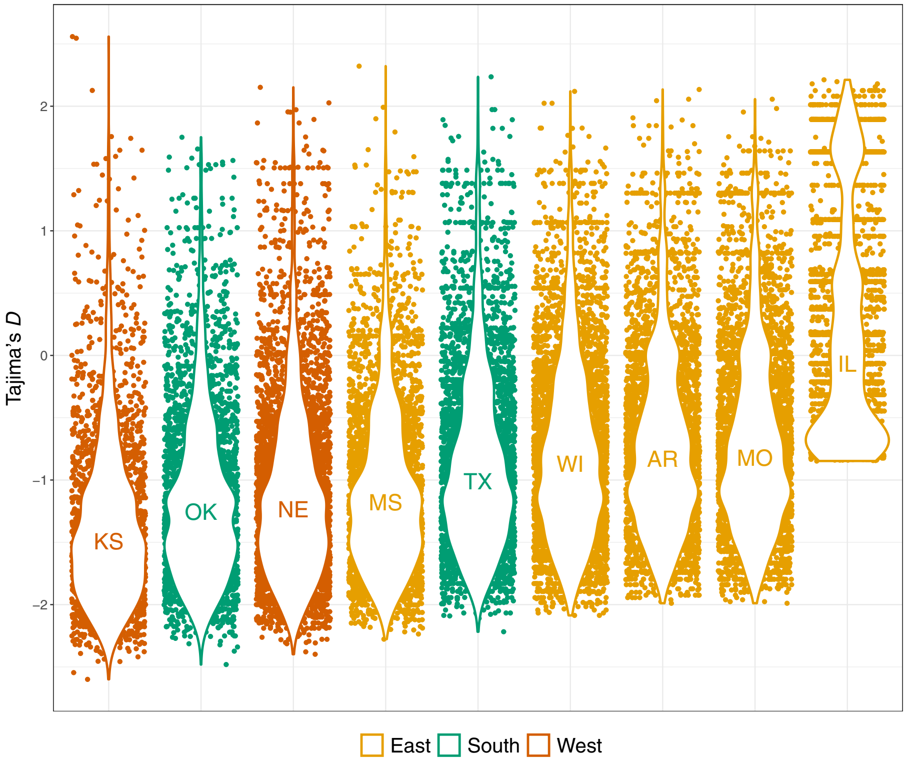

FIGURE 5. Distribution of Tajima's *D* across contigs among grouped plants within each sampling state of origin. Colors correspond to different regional groupings inferred from PCA, conStruct, and SVDQuartets analysis. Across markers, the majority of Tajima's *D* estimates are negative, consistent with a recent species-wide range expansion. We hypothesize that the elevated Tajima's *D* found in Illinois samples are the result of recent introgression between Illinois and non-eastern region plants.

BayEnv2. We speculated that the combined knowledge of the protein function from BLAST and eggNOG searches and associated environmental variables would lead to better understanding of past selective pressures *S. integrifolium* experienced while expanding to its current geographic range. Several genes with known functions associated with abiotic stress tolerance show associations with climatic variables. Below, we highlight environmental variables that are associated with genes to further understand biological processes that may have enabled range expansion. For example, Δ8-fatty-acid desaturase has a role in tolerating cold and chilling environments (Chen et al., 2012) and was found to be associated with both continentality (temperature difference between warmest and coldest months) and mean annual precipitation variables. Aquaporin TP1 encodes a channel protein involved in water transport that has been shown to influence cell turgor and expansion (Maurel et al., 1993) and was found to be associated with annual heat moisture index, a composite measure of heat and precipitation, and longitude. The E3 ubiquitin ligase gene *SUD1* is involved in drought tolerance through its effects on cuticle wax formation and transpiration rate (Lü et al., 2012) and was found to be associated with the mean warmest month temperature and extreme warm temperatures.

TABLE 2. Results from f_3 tests for introgression to test the hypothesis that the Illinois population is admixed. Significantly negative test statistics indicate that Illinois population has history of a non tree-like branching process, i.e., past admixture between populations P1 and P2.

Focal	P1	P2	f3	SE	*Z*-score	*p*-value
Illinois	Wisconsin	Mississippi	−0.0012	0.0010	−1.1329	1.286e-01
Illinois	Wisconsin	Arkansas	−0.0004	0.0011	−0.3531	3.620e-01
Illinois	Wisconsin	Missouri	−0.0004	0.0010	−0.3724	3.548e-01
Illinois	Wisconsin	Texas	−0.0031	0.0011	−2.7611	2.881e-03
Illinois	Wisconsin	Oklahoma	−0.0042	0.0012	−3.5755	1.748e-04
Illinois	Wisconsin	Kansas	−0.0045	0.0010	−4.3112	8.118e-06
Illinois	Wisconsin	Nebraska	−0.0044	0.0011	−4.1234	1.867e-05

DISCUSSION

Reference transcriptome as a resource

The development of *Silphium* genomic resources has been hindered because of difficulties imposed by its large genome that is replete with repetitive elements. We sought transcriptome resequencing as a reliable method to obtain a common set of reduced representation genetic markers from this nonmodel organism. Toward this goal, we built a reference transcriptome of a diverse set of tissues from a single plant and used it as a genotyping reference for the resequencing of a species-wide panel of *S. integrifolium* samples. The initial transcriptome assembly contained roughly 250 Mb of sequence data assembled into more than 166,000 contigs, some of which are products of sequence contamination and assembly error. An assembly of the longest isoform of each contig contained a majority of nearly full-length transcripts and an almost complete set of universal benchmark genes predicted to be present in all plants. Our final genotyping reference was the product of filtering to include only contigs that contained genes that have been sequenced in other plants and present as single copies in both *Silphium* and its close relative, *Helianthus annuus*. Assuming approximate conservation of the gene number present across plants, the reference transcriptome contains roughly one third of genes in *Silphium* and can presumably serve as a helpful genotyping resource in place of a quality reference genome (Sterck et al., 2007).

S. integrifolium comprises three genetic groups

We resequenced the young leaf transcriptomes from dozens of *S. integrifolium* plants from localities that cover the species range. Multiple analyses to identify genetic population structure across samples found that *S. integrifolium* comprises three clusters (Figs. 1, 2). Although the genetic clusters are each found within particular geographic regions of the United States, there are no obvious physical barriers that may help maintain cluster integrity by restricting gene flow. Groups of samples that are geographically near (≈ 200 km) but belong to different regional groupings show greater differentiation, as measured by F_{ST}, than the most distant (~ 1000 km)

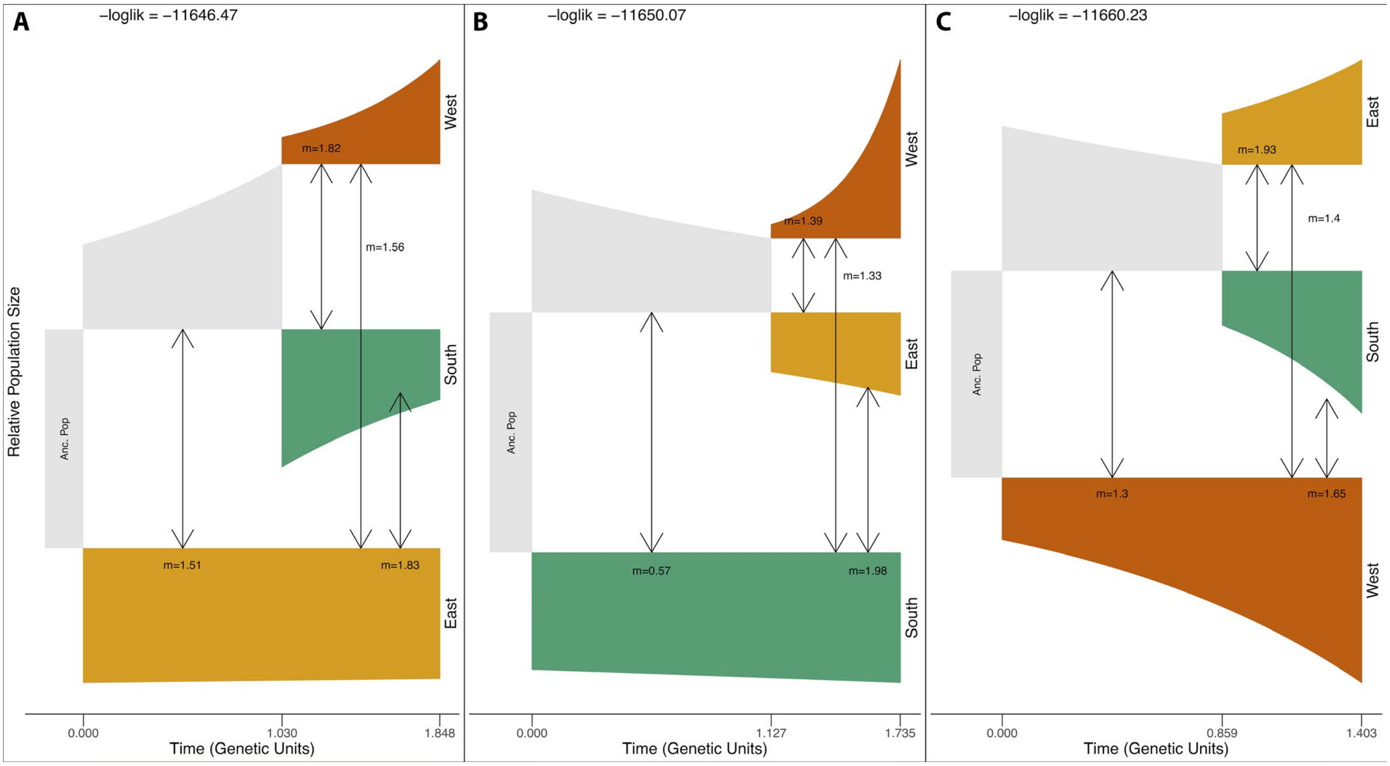

FIGURE 6. Summarized results of best-fitting demographic inference model (symmetric migration rates) for each lineage branching scenario, arranged left to right from highest to lowest log-likelihood. For each plot, sampled extant lineages are colored, and ancestral lineages are gray. Each simulation evolves from an initial ancestral population (leftmost gray box) and undergoes two splitting events to produce three extant lineages (right edge of each plot). The log-likelihood is reported above each plot. Height of each lineage segment is equal to population size relative to ancestral population (Anc. Pop) size. Migration rate estimates are reported near arrows in each figure. Estimated model parameters with standard deviations and model fit statistics are listed in Table 3. Time and migration rates are in genetic units ($2N_e$).

pairs from the same regional grouping (Fig. 3). Although we found considerable levels of neutral genetic variation within each population or region, both are lower than levels found in other self-incompatible wild sunflower species (Fig. 4). Using a set of nuclear genes, (Liu and Burke, 2006) found approximately three times the amount of synonymous variation in wild *H. annuus* (wild populations $\pi = 0.0315$) than what we observe in *S. integrifolium*. The discrepancy suggests that *S. integrifolium* experienced a much smaller historical effective population size than the annual sunflower, perhaps reflecting the longer generation time of *S. integrifolium* as a long-lived perennial and/or other features of its life history (Andreasen and Baldwin, 2001).

Plants from the eastern region form the most genetically variable cluster (Fig. 4B). Within the eastern region, there is a general negative correlation between latitude and π (Fig. 4A). The Illinois samples were found to have appreciable amounts of admixed ancestry (Fig. 2A, B). Results from f_3 tests for introgression having a role in the Illinois population's past were significant (Table 2). Introgression from the west could potentially have proceeded naturally through unsampled populations, e.g., Iowa, or from western *S. integrifolium* seed introduced by humans in prairie restoration projects. Within the eastern region, populations are separated by latitude in phylogenetic analyses (Fig. 2C).

The southern region maintains a breadth of genetic variation that is comparable to the variation of the eastern region (Fig. 4B). There is a negative correlation between latitude and genetic diversity across the two states. All southern region samples were found to be well described by a mix of its own and other ancestry (Fig. 2B). The Texas samples are best described as a mixture of ancestry from the South and Eastern regions and the Oklahoma samples as a mixture of ancestry from the South and Eastern regions. We note that this description does not necessarily imply recent gene flow or admixture.

The western region contains the least amount of nucleotide variation (Fig. 4B). The same negative correlation between latitude and genetic variation is observed in western samples (Fig. 4A). Almost all western samples were inferred to be overwhelmingly made up of a single ancestry group (Fig. 2B). The reduced levels of polymorphism in the west result in strong phylogenetic clustering that largely groups with geography. Kansas samples are paraphyletic with Nebraska samples forming a clade that is embedded within the larger western clade. All western samples form a clade that is sister to a sample from the Oklahoma population (Fig. 2C).

Geographic origin of *S. integrifolium*

It has been claimed that *S. integrifolium* originated from within the historical range of the American tallgrass prairie, i.e., the prairie origin hypothesis (Settle and Fisher, 1972). The close association of the species with the prairie habitat and the inferred young age of the species were cited as evidence for this assertion. More recently, a phylogenetic treatment of the *Silphium* genus reconstructed

TABLE 3. Demographic inference model parameter estimates with standard deviations. For each scenario topology, e.g., (0, (1, 2)), *s* parameters are the proportion of lineage that split into population (s0 = proportion of ancestral population that split into population 0), nu parameters are population sizes, *m* parameters are migration rates between listed populations, T1 is time since initial split, T2 is time since second split.

Topology	AIC	Loglik	s0	s1	nu12	nu0	nu1	nu2	m0_12	m0_2	m0_1	m1_2	T1	T2
(East, (South, West))	23,316.94	−11,646.94	0.614 (0.503)	0.837 (0.022)	0.753 (1.865)	0.596 (1.06)	0.323 (0.378)	0.476 (0.079)	1.507 (11.426)	1.557 (0.38)	1.833 (4.908)	1.816 (0.867)	1.848 (3.91)	0.818 (2.712)
(South, (East, West))	23,324.14	−11650.07	0.489 (0.092)	0.805 (0.211)	0.308 (0.202)	0.543 (0.097)	0.347 (0.258)	0.744 (0.662)	0.573 (0.74)	1.326 (1.459)	1.982 (0.036)	1.386 (0.841)	1.735 (0.094)	0.609 (0.027)
(West, (South, East))	23,344.46	−11,660.23	0.3 (1.054)	0.516 (0.005)	0.513 (0.332)	0.991 (1.011)	0.688 (0.333)	0.507 (0.233)	1.303 (3.305)	1.397 (0.041)	1.654 (0.651)	1.925 (1.128)	1.403 (6.349)	0.544 (0.027)

S. integrifolium to be most closely related to the common prairie cup plant *S. perfoliatum*, and *S. wasiotense* (Clevinger and Panero, 2000). Medley (1989) used populations endemic to the dry-mesic, mixed hardwood forests of eastern Kentucky to describe *S. wasiotense* and estimated the species to be closely related to other congeners found on limestone and sandstone soils throughout the unglaciated Cumberland Plateau (*S. brachiatum* and *S. mohrii*). The present study represents the first application of population genetic data to the topic of identifying the geographic origin of *S. integrifolium*. Our results contradict the prairie origin hypothesis and suggest that the geographic origin of *S. integrifolium* is in the American Southeast.

We hypothesize that *S. integrifolium* has expanded its range from unglaciated regions near present day Mississippi or Texas to cover the American prairie, i.e., refugium origin hypothesis (see Fig. 1A for species range estimate; Appendix S8). The evidence for our refugium origin hypothesis is multifold. First, there is a consistent correlation between latitude and diversity. Mississippi plants are the most genetically diverse samples. Texas plants, from the southern region, are also very diverse. However, the level of divergence found between the two groups of plants is within the range of variation found within Mississippi, but not Texas. From the sequence divergence data, this pattern suggests that Texas plants are derived from a population somewhat like our Mississippi sample; however, more sampling is required to better understand the progenitor taxon in this split. Results from demographic modeling using three-dimensional joint site frequency spectra generally support the refugium origin hypothesis (Fig. 6); however, confidence intervals around exact parameter estimates are broad (Table 3). We hypothesize that after the split between Mississippi and Texas populations, both eastern and southern lineages expanded northward, reducing genetic variation with northward progression. The expansion of Oklahoma populations into Kansas and Nebraska resulted in a large reduction in genetic diversity. This reduction in variation appears to be most recent and results in western samples still having considerable amounts of the ancestry layer that is dominant in southern samples (Fig. 2A, B).

The overall negative values of Tajima's *D* in all groups, excluding Illinois, suggest a relatively recent population expansion for the entire species. We interpret the patterns of Tajima's *D* across contigs to help us identify the most recent population expansion. Western and southern regions have the largest negative values of Tajima's *D*, indicating a large excess of rare variants possibly due to a very recent population expansion after bottlenecks. The smaller negative Tajima's *D* values found in the other eastern region may be the result of a more modest and/or a more ancient expansion, after which genetic variation has started to accumulate at sites with initially rare variants.

The results of our phylogenetic analyses support the refugium origin hypothesis. We observe a series of nested relationships within the SVDQuartets phylogeny (Fig. 2C). Western samples form a clade that is nested within a larger clade also containing southern samples. The larger clade containing all southern and western samples is sister to an eastern sample clade. On the basis of pubescence and number of ray florets of our specimens (unpublished data), we suggest that the eastern clade corresponds to the subspecies *S. integrifolium* var. *integrifolium* and the other clade corresponds to the subspecies *S. integrifolium* var. *laeve* (http://www.efloras.org). The high levels of observed variation in Mississippi are consistent with samples being placed throughout the eastern region clade and outside of the larger

S. integrifolium clade. The results of our TWISST analyses suggest that the general topological pattern we observe in our SVDQuartets phylogeny has support across the majority of transcriptome markers. The southern and western regions are found to be sister taxa to each other more often than either is to the eastern region taxon. This pattern is consistent with our demographic inference results that western and southern regions have diverged most recently within this species and that their common ancestor split from an eastern region population (Fig. 5). Southern and western regions are found to be sister to the eastern region in equal proportions. This pattern is expected under the refugium origin hypothesis.

The two competing hypotheses are not necessarily mutually exclusive. The prairie origin hypothesis (Settle and Fisher, 1970, 1972) coupled with the historical range of the prairie (Kurtz, 2013) implicitly suggested to many that the central prairie was the center of *S. integrifolium* diversity. At its largest expanse, the historical prairie included a small amount of habitat in present day Mississippi that was geographically separated from the larger contiguous prairie (Kurtz, 2013). We cannot rule out that ancestral *S. integrifolium* originated in the small, restricted prairie habitat of Mississippi. However, an origin in the geographic center of the current prairie is not consistent with the genetic variation observed. The best remaining hypotheses, given our observed patterns of genetic variation, would be an origin in the southeastern edge of its current range. A more thorough sampling of Mississippi and Texas *S. integrifolium* and wide geographic sampling of *S. wasiotense*, and *S. perfoliatum* are recommended to further investigate the geographic origin of *S. integrifolium* and interspecific relationships.

Potentially adaptive candidate genes

Outlier analyses based on BayeScan indicate that several dozens of the genes with the highest average divergence have SNPs that are statistical outliers, falling outside the distribution of most transcriptome markers. Several of these genes show interesting annotations. Together, these candidate loci may have been involved in adaptation to the drier environment with more extreme temperatures experienced during the hypothesized expansion of *S. integrifolium* out of the American Southeast.

The SNPs with extremely low false discovery rate ($q < 0.001$) have patterns of genetic variation highly unlikely to have occurred by chance. The best explanation for these patterns is that they are very tightly linked to genetic variation under selection during local adaptation, and indeed may be in genes important for adaptation. These genes encode proteins involved in transcriptional regulation (transcriptional corepressor SEUSS), secondary compound metabolism, disease tolerance, and abiotic stress responses (Δ8-fatty-acid desaturase, aquaporin TP1), all of which could have been important for adapting to local conditions. To better understand the local environment facets that were potential drivers of adaptation, we looked for climatic variables that were correlated with each selected protein across the landscape. The large number of mitochondrial and chloroplast-targeted proteins suggests a possible role for cytonuclear interactions. The presence of several genes involved in cytokinin responses may be explained by altered root to shoot ratios being important in developing the extreme drought tolerance for which this species is known. In addition to being candidates that may underlie local adaptation, many of these genes with high divergence make them interesting to examine further as targets for breeding, particularly in cases where they may alter phenotypes important for domestication. As an example, the variation found in 13S globulin seed storage protein 2 is potentially important because in other species this protein family is a major driver of seed nutritional content.

Including diversity into breeding program

Perhaps motivated by the prairie origin hypothesis, *S. integrifolium* breeding programs overwhelmingly have used germplasm derived from collections from the western region (Vilela et al., 2018). From those limited collections, significant heritable domestication improvements have been made in the accumulation of aboveground biomass and seed yield (Vilela et al., 2018). Our population genetic results suggest that past domestication improvements have been achieved with a limited genetic variation resource. The western region, while containing considerable genetic variation, is depauperate of variation relative to the eastern and southern regions. Evidence for population recovery from a recent genetic bottleneck (e.g., low Tajima's *D* values, Fig. 5) could imply that potentially useful variation may have been lost during adaptation to this drought prone climate.

The elevated genetic variation across the rest of the species range and the increased divergence between regions is evidence that useful variation that may be beneficial to domestication efforts can be further incorporated into the breeding program. Characterizing the potential primary gene pool of a crop can help breeders identify useful parents in at least two ways. First, demonstrating trait or allele association with climatic, soil, or geographic variables could enable breeders to predict which wild populations might produce individuals with extreme values for traits of interest. Recent studies have found that the domestication-targeted traits, seed size, and seed oil profile in wild populations have clear geographic patterns (Reinert et al., 2019). Second, information about the distribution of genetic variation within and between populations and phylogenetic relationships between populations may inform the sampling strategy for a wild germplasm collection campaign designed to capture the most genetic variation in the least number of accessions that need to be subsequently characterized and maintained. The kind of genetic variation thus preserved ex situ may require careful screening (e.g., for resistance to a particular pathogen) to be valuable to breeders or may be held in reserve for future breeders seeking new genetic variation to adapt crops to evolving pests or climate change. This information should also help inform in situ conservation strategies, and in this case, it argues for high priority to be placed on conserving rare native prairies in the southern United States (Campbell and Seymour, 2011) as these appear to be the center of origin of at least one *Silphium* species and perhaps a similar pattern will be found for other native grassland species.

ACKNOWLEDGMENTS

The authors thank the Associate Editor and anonymous reviewers for their helpful comments and suggestions that improved the manuscript; John Holmquist, who found and collected many of the wild populations and helped with RNA extractions; other members of the public, botanical gardens, and prairie conservation organizations who have donated wild *Silphium* germplasm to the The Land Institute's collection over the years; the financial support of the Perennial Agriculture Project and The Land Institute's donors; Kelsey Peterson, John Hill Price, and Kevin P. Smith for their discussion and advice which improved the manuscript; and the National Science Foundation for grant #1737827 Dimensions US-China to Y.B.

AUTHOR CONTRIBUTIONS

D.V.T., N.K., and Y.B. conceptualized the study. A.R., A.H., C.P., and K.D. performed formal analysis and data curation. Y.B. and D.V.T. acquired funding for the project. A.R. wrote the first draft of the manuscript. All authors contributed to the review and editing of subsequent drafts and approved the final version of the manuscript.

DATA AVAILABILITY

Transcriptome resequencing alignment and genotyping reference data files have been deposited in the NCBI Sequence Read Archive (BioProjectID PRJNA665243, BioSample accessions SAMN16247094–SAMN16247161).

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

APPENDIX S1. Geographic locality for all *Silphium integrifolium* samples used in transcriptome resequencing study.

APPENDIX S2. Description of climatic variables used in BayEnv2 analyses of genotype–environment associations.

APPENDIX S3. Summary of TWISST support for the SVDQuartets inferred phylogeny topology.

APPENDIX S4. Summarized nucleotide variation found across *Silphium integrifolium* samples.

APPENDIX S5. Associations between BayeScan inferred divergence outlier proteins and environmental variables sorted by highest quantile Bayes Factor.

APPENDIX S6. Associations between BayeScan inferred divergence outlier proteins and environmental variables sorted by highest quantile of absolute Pearson correlation coefficient.

APPENDIX S7. Associations between BayeScan inferred divergence outlier proteins and environmental variables sorted by highest quantile of absolute Spearman's rank correlation coefficient.

APPENDIX S8. Methods for MaxEnt species distribution modeling.

LITERATURE CITED

Andreasen, K., and B. G. Baldwin. 2001. Unequal evolutionary rates between annual and perennial lineages of checker mallows (*Sidalcea*, Malvaceae): evidence from 18S–26S rDNA internal and external transcribed spacers. *Molecular Biology and Evolution* 18: 936–944.

Bai, C., W. S. Alverson, A. Follansbee, and D. M. Waller. 2012. New reports of nuclear DNA content for 407 vascular plant taxa from the United States. *Annals of Botany* 110: 1623–1629.

Boutet, E., D. Lieberherr, M. Tognolli, M. Schneider, P. Bansal, A. J. Bridge, S. Poux, et al. 2016. UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: How to use the entry view. *In* D. Edwards [ed.], Plant bioinformatics, Methods in molecular biology, vol. 1374, 23–54.Humana Press, New York, NY, USA.

Bradburd, G. S., G. M. Coop, and P. L. Ralph. 2018. Inferring continuous and discrete population genetic structure across space. *Genetics* 210: 33–52.

Broad Institute. 2018. Picard version 2.18.16. Available at http://broadinstitute.github.io/picard/. Broad Institute, Cambridge, MA, USA.

Bryant, D. M., K. Johnson, T. DiTommaso, T. Tickle, M. B. Couger, D. Payzin-Dogru, T. J. Lee, et al. 2017. A tissue-mapped Axolotl de novo transcriptome enables identification of limb regeneration factors. *Cell Reports* 18: 762–776.

Buffalo, V. S. 2014. Scythe – A Bayesian adapter trimmer. Website: https://github.com/vsbuffalo/scythe [accessed 01 June 2019].

Bushnell, B. 2018. Bbtools: A suite of fast, multithreaded bioinformatics tools designed for 493 analysis of DNA and RNA sequence data. Website: https://jgi.doe.gov/494data-and-tools/bbtools495. DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.

Campbell, J. J. N., and W. R. Seymour. 2011. The vegetation of Pulliam Prairie, Chickasaw County, Mississippi: a significant remnant of pre-Columbian landscape in the Black Belt. *Journal of the Mississippi Academy of Sciences* 56: 248–263.

Charlesworth, D. 2003. Effects of inbreeding on the genetic diversity of populations. *Philosophical Transactions of the Royal Society, B, Biological Sciences* 358: 1051–1070.

Chen, M., J. E. Markham, and E. B. Cahoon. 2012. Sphingolipid Δ 8 unsaturation is important for glucosylceramide biosynthesis and low-temperature performance in Arabidopsis. *Plant Journal* 69: 769–781.

Chifman, J., and L. Kubatko. 2014. Quartet inference from SNP data under the coalescent model. *Bioinformatics* 30: 3317–3324.

Clevinger, J. A., and J. L. Panero. 2000. Phylogenetic analysis of *Silphium* and subtribe Engelmanniinae (Asteraceae: Heliantheae) based on ITS and ETS sequence data. *American Journal of Botany* 87: 565–572.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, et al. 2011. The Variant Call Format and VCFtools. *Bioinformatics* 27: 2156–2158.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, et al. 2013. STAR: ultrafast universal RNA-Seq aligner. *Bioinformatics* 29: 15–21.

Emms, D. M., and S. Kelly. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* 16: 157.

Fiedler, A. K., and D. A. Landis. 2007. Attractiveness of Michigan native plants to arthropod natural enemies and herbivores. *Environmental Entomology* 36: 751–765.

Foll, M. 2012. BayeScan V2. 1 user manual. *Ecology* 20: 1450–1462.

Foll, M., and O. Gaggiotti. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* 180: 977–993.

Fox, D. L., and P. L. Koch. 2003. Tertiary history of C_4 biomass in the Great Plains, USA. *Geology* 31: 809–812.

Fox, D. L., and P. L. Koch. 2004. Carbon and oxygen isotopic variability in neogene paleosol carbonates: constraints on the evolution of the C_4-grasslands of the Great Plains, USA. *Palaeogeography, Palaeoclimatology, Palaeoecology* 207: 305–329.

Garrison, E., and G. Marth. 2012. Haplotype-based variant detection from short-read sequencing. arXiv 1207.3907 [Preprint].

Günther, T., and G. Coop. 2013. Robust identification of local adaptation from allele frequencies. *Genetics* 195: 205–220.

Haas, B., and A. Papanicolaou. 2017. TransDecoder. https://TransDecoder.github.io [accessed 15 July 2019].

Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, et al. 2013. De novo transcript sequence reconstruction from RNA-Seq using the Trinity platform for reference generation and analysis. *Nature Protocols* 8: 1494.

Isaacs, R., J. Tuell, A. Fiedler, M. Gardiner, and D. Landis. 2009. Maximizing arthropod-mediated ecosystem services in agricultural landscapes: the role of native plants. *Frontiers in Ecology and the Environment* 7: 196–203.

Jouganous, J., W. Long, A. P. Ragsdale, and S. Gravel. 2017. Inferring the joint demographic history of multiple populations: beyond the diffusion approximation. *Genetics* 206: 1549–1567.

Knapp, A. K., J. M. Briggs, S. L. Collins, S. R. Archer, M. S. Bret-Harte, B. E. Ewers, D. P. Peters, et al. 2008. Shrub encroachment in North American grasslands: shifts in growth form dominance rapidly alters control of ecosystem carbon inputs. *Global Change Biology* 14: 615–623.

Kurtz, C. 2013. A practical guide to prairie reconstruction. University of Iowa Press, Ames, IA, USA.

Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9: 357.

Lauenroth, W. K. 1979. Grassland primary production: North American grasslands in perspective. *In* N. R. French [ed.], Perspectives in grassland ecology, 3–24.Springer, New York, NY, USA.

Leopold, A. 1968. A Sand County almanac, and sketches here and there. Oxford University Press, New York, NY, USA.

Li, B., and C. N. Dewey. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12: 323.

Liu, A., and J. M. Burke. 2006. Patterns of nucleotide diversity in wild and cultivated sunflower. *Genetics* 173: 321–330.

Lü, S., H. Zhao, D. L. Des Marais, E. P. Parsons, X. Wen, X. Xu, D. K. Bangarusamy, et al. 2012. Arabidopsis *ECERIFERUM9* involvement in cuticle formation and maintenance of plant water status. *Plant Physiology* 159: 930–944.

Madden, T. 2013. The BLAST sequence analysis tool. *In* The NCBI handbook, 2nd ed. [online]. Website: https://www.ncbi.nlm.nih.gov/books/NBK143764/ National Center for Biotechnology Information, Bethesda, MD, USA.

Martin, S. H., and S. M. Van Belleghem. 2017. Exploring evolutionary relationships across the genome using topology weighting. *Genetics* 206: 429–438.

Matthews, W. J. 1988. North American prairie streams as systems for ecological study. *Journal of the North American Benthological Society* 7: 387–409.

Maurel, C., J. Reizer, J. I. Schroeder, and M. J. Chrispeels. 1993. The vacuolar membrane protein gamma-TIP creates water specific channels in *Xenopus* oocytes. *EMBO Journal* 12: 2241–2247.

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

Medley, M. 1989. *Silphium wasiotensis* (Asteraceae), a new species from the Appalachian Plateaus in eastern Kentucky. *SIDA, Contributions to Botany* 13: 285–291.

Nei, M. 1972. Genetic distance between populations. *American Naturalist* 106: 283–292.

Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York, NY, USA.

Paradis, E. 2010. Pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics* 26: 419–420.

Pembleton, L. W., N. O. I. Cogan, and J. W. Forster. 2013. StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Molecular Ecology Resources* 13: 946–952.

Pickrell, J. K., and J. K. Pritchard. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics* 8: e1002967.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959.

Reich, D., K. Thangaraj, N. Patterson, A. L. Price, and L. Singh. 2009. Reconstructing Indian population history. *Nature* 461: 489–494.

Reinert, S., D. L. Van Tassel, B. Schlautman, N. C. Kane, and B. S. Hulke. 2019. Assessment of the biogeographical variation of seed size and seed oil traits in wild *Silphium integrifolium* Michx. genotypes. *Plant Genetic Resources: Characterization and Utilization* 17: 427–436.

Samson, F. B., F. L. Knopf, and W. R. Ostlie. 2004. Great Plains ecosystems: past, present, and future. *Wildlife Society Bulletin* 32: 6–14.

Samson, F., and F. Knopf. 1994. Prairie conservation in North America. *BioScience* 44: 418–421.

Sayers, E. W., T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, and V. Chetvernin, et al. 2009. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 38(Database issue): D5–D16.

Settle, W. J., and T. R. Fisher. 1972. A biosystematic study of *Silphium integrifolium* Michaux (Compositae). *Ohio Journal of Science* 72: 254–265.

Settle, W. J., and T. R. Fisher. 1970. The varieties of *Silphium integrifolium*. *Rhodora* 72: 536–543.

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31: 3210–3212.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.

Stanford, G. 1990. *Silphium perfoliatum* (cup plant) as a new forage. *In* D. D. Smith and C. A. Jacobs [eds.], Proceedings of the Twelfth North American Prairie Conference, 33–37, Cedar Falls, IA, USA.

Sterck, L., S. Rombauts, K. Vandepoele, P. Rouze, and Y. Van de Peer. 2007. How many genes are there in plants (… and why are they there)? *Current Opinion in Plant Biology* 10: 199–203.

Swofford, D. L. 2003. PAUP*: phylogenetic analysis using parsimony (*and other methods), version 4.0. Sinauer, Sunderland, MA, USA.

Ustak, S., and J. Munoz. 2018. Cup-plant potential for biogas production compared to reference maize in relation to the balance needs of nutrients and some microelements for their cultivation. *Journal of Environmental Management* 228: 260–266.

Van Tassel, D. L., K. A. Albrecht, J. D. Bever, A. A. Boe, Y. Brandvain, T. E. Crews, M. Gansberger, et al. 2017. Accelerating *Silphium* domestication: an opportunity to develop new crop ideotypes and breeding strategies informed by multiple disciplines. *Crop Science* 57: 1274–1284.

Vilela, A., L. González-Paleo, K. Turner, K. Peterson, D. Ravetta, T. Crews, and D. Van Tassel. 2018. Progress and bottlenecks in the early domestication of the perennial oilseed *Silphium integrifolium*, a sunflower substitute. *Sustainability* 10: 638.

Wang, C., Z. A. Szpiech, J. H. Degnan, M. Jakobsson, T. J. Pemberton, J. A. Hardy, A. B. Singleton, and N. A. Rosenberg. 2010. Comparing spatial maps of human population-genetic variation using Procrustes analysis. *Statistical Applications in Genetics and Molecular Biology* 9: 13.

Wang, T., A. Hamann, D. Spittlehouse, and C. Carroll. 2016. Locally downscaled and spatially customizable climate data for historical and future periods for North America. *PLoS One* 11: e0156720.

Weaver, J. E. 1954. North American Prairie. Johnsen Publishing, Lincoln, NE, USA.

Weaver, J. E., L. A. Stoddart, and W. M. Noll. 1935. Response of the prairie to the great drought of 1934. *Ecology* 16: 612–629.

Weir, B. S., and C. C. Cockerham. 1984. Estimating *F*-statistics for the analysis of population structure. *Evolution* 38: 1358–1370.

Wever, C., M. Höller, L. Becker, A. Biertümpfel, J. Köhler, D. van Inghelandt, P. Westhoff, et al. 2019. Towards high-biomass yielding bioenergy crop *Silphium perfoliatum* L.: phenotypic and genotypic evaluation of five cultivated populations. *Biomass and Bioenergy* 124: 102–113.

Williams, J. W., B. N. Shuman, T. Webb III, P. J. Bartlein, and P. L. Leduc. 2004. Late-Quaternary vegetation dynamics in North America: scaling from taxa to biomes. *Ecological Monographs* 74: 309–334.

Wilson, G. W. T., and D. C. Hartnett. 1998. Interspecific variation in plant responses to mycorrhizal colonization in tallgrass prairie. *American Journal of Botany* 85: 1732–1738.