Haplotype-phased genomes elucidate genetic architecture of environmental adaptation and domestication traits in Silphium

David Van Tassel

vantassel@landinstitute.org

The Land Institute https://orcid.org/0000-0003-0844-3976

Renan Souza

HudsonAlpha Institute for Biotechnology

Josh Clevenger

HudsonAlpha Institute for Biotechnology https://orcid.org/0000-0002-9830-9481

Jerry Jenkins

HudsonAlpha Institute for Biotechnology https://orcid.org/0000-0002-7943-3997

Walid Korani

HudsonAlpha Institute for Biotechnology

Jeremy Schmutz

HudsonAlpha Institute for Biotechnology https://orcid.org/0000-0001-8062-9172

Jane Grimwood

HudsonAlpha Institute for Biotechnology https://orcid.org/0000-0002-8356-8325

Chris Plott

HudsonAlpha Institute for Biotechnology https://orcid.org/0000-0002-0109-5174

Yaniv Brandvain

University of Minnesota

Kathryn Turner

The Land Institute

Allison Miller

Donald Danforth Plant Science Center

Brent Hulke

Agricultural Research Service https://orcid.org/0000-0001-5380-0827

Louis Nastasi

University of Iowa https://orcid.org/0000-0001-7825-480X

Ebony Murrell

Savanna Institute

Lori Handley

HudsonAlpha Institute for Biotechnology
Jenell Weber
Genome Sequencing Center, HudsonAlpha Institute for Biotechnology
Lori Boston
Genome Sequencing Center, HudsonAlpha Institute for Biotechnology
Stella Woeltjen
Donald Danforth Plant Science Center
Matthew Rubin
Saint Louis University
Betsy Trana
The Land Institute
Erin Howard
University of Georgia
Ashlyn Bazzel
Auburn University https://orcid.org/0009-0002-8391-9372
Alex Harkess
HudsonAlpha Institute for Biotechnology https://orcid.org/0000-0002-2035-0871

Article

Keywords:

Posted Date: February 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8753360/v1

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Nature Communications on July 9th, 2026. See the published version at https://doi.org/10.1038/s41467-026-75205-3.

Haplotype-phased genomes elucidate genetic architecture of environmental adaptation and domestication traits in *Silphium*

Renan Souza[1], Josh Clevenger[1], Jerry Jenkins[1], Walid Korani[1], Jeremy Schmutz[1], Jane Grimwood[1], Christopher Plott[1], Yaniv Brandvain[2], Kathryn Turner[3], Allison J. Miller[4,5], Brent S. Hulke[6], Louis F. Nastasi[7], Ebony G. Murrell[8], Lori H. Handley[1], Jenell Webber[1], Lori Beth Boston[1], Stella Woeltjen[4], Matthew J. Rubin[5], Betsy Trana[3], Erin Howard[9], Ashlyn Bazzel[10], Alex Harkess[1], David L. Van Tassel[3*]

Affiliations

[1] HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA

[2] University of Minnesota, Minneapolis, MN, USA

[3] The Land Institute, Salina, KS, USA

[4] The Donald Danforth Plant Science Center, Saint Louis, MO, USA

[5] Saint Louis University, Saint Louis, MO, USA

[6] USDA-Agricultural Research Service, Edward T Schafer Agricultural Research Center, Fargo, ND, USA

[7] Department of Biology, University of Iowa, Iowa City, IA, USA

[8] Savanna Institute, Madison, WI, USA

[9] The University of Georgia, Athens, GA, USA

[10] Auburn University, Auburn, AL, USA

*Corresponding author: vantassel@landinstitute.org

Abstract

Developing native perennial crops is vital for climate-resilient agriculture, yet their domestication is often hindered by a lack of cost-effective genomic resources. To build a framework for genomics-accelerated domestication of perennials with large, complex genomes, we generated chromosome-scale, haplotype-phased assemblies for *Silphium integrifolium* Michx. and *Silphium perfoliatum* L., two deep-rooted native North American prairie species valued for drought tolerance and oil production. The genomes are characterized by the presence of a putative helical structure preserved during interphase with a loop circumference of 43 Mb. Using targeted sequencing of 14 *Silphium* species, we first refined the phylogeny of the genus, recovering the *Composita* and *Silphium* subgenera while identifying taxonomic discrepancies. We used a combination of low-coverage short-read sequencing and target-sequencing to characterize a 258-accession diversity panel to define ancestral populations and perform genome-wide association studies (GWAS) on several traits. We identified 81 loci associated with environmental adaptation and domestication traits; notably, variants in a *MATE* transporter, Alpha-Beta hydrolase and a *ACR4*-related protein explained significant variance in seed number and floral architecture. These findings establish the genomic framework necessary to accelerate the domestication of *Silphium* and provide a model for unlocking the potential of other complex wild genomes.

1 INTRODUCTION

Global food security rests on a precarious foundation. Over 50% of human caloric intake is estimated to be derived from just five crop species: rice, wheat, maize, sugarcane, and barley[1]. Our dependence on just a few plant species for global survival poses serious food security concerns due to risks of supply disruption amid geopolitical issues, climate change, and the growing need to increase food production to sustain the human population. One part of the solution to this problem lies in domestication and breeding of the nearly 30,000 wild but edible plant species that are not actively cultivated[2,3]. Among these wild candidates, perennial species hold particular promise and value, as they may provide an avenue to marry high productivity, lower agricultural inputs, and ecosystem sustainability[4].

Traditional domestication paths took millennia and their strong selection for germination, increased yield and harvestability in specific environments, have led to the unintended loss of favorable alleles for other traits such as nutritional value and stress resistance[5]. Modern development of novel crops from wild plants has been rare, but such "*de novo* domestication" efforts should be greatly accelerated by recent advances in genomics and breeding methods[6]. Genomic methods and technologies such as reference genome assembly, large-scale genotyping, genome wide association studies, genomic prediction and gene editing accelerate the improvement of agronomic performance and, simultaneously, the retention of important traits such as biotic and abiotic stress resistance, resulting in more resilient crops. Therefore, *de novo* domestication can unlock the potential of wild plant species suitable for human use because it is both more precise and faster than the original domestication pathways.

One genus with great potential for multiple agricultural uses is *Silphium* L. (Asteraceae: Heliantheae), a native of the North American prairies. The genus includes around 20 diploid (2n = 14) species[7], most of which reproduce by outcrossing, including interspecific hybridization (Reinert et al., 2020). *Silphium integrifolium* Michx. and *S. perfoliatum* L. are currently undergoing domestication, with the first showing promise as a new oilseed crop with high tolerance to drought[8,9] and the second as a biomass and fiber crop[10]. *Silphium integrifolium* seeds have considerable concentration of oil (11.8 to 25.3 % w/w; Reinert et al., 2019) and high concentrations of squalene (4.9% of the oil), a triterpene used as an emollient in cosmetics and as an adjuvant in vaccines [9]. Squalene has traditionally been sourced from shark livers, but in light of shark conservation efforts, producing this compound from plants is of great interest[11].

In addition to their potential as oilseed and biomass crops, *Silphium* species play key ecological roles in the prairie ecosystem[12]. This iconic group of plants is critical for prairie biodiversity maintenance and can be strategically reintroduced to disturbed sites through

ecological restoration. As a perennial plant with a deep root structure, *Silphium* can maintain soil integrity, build soil carbon and produce seeds even during long drought periods. The *Silphium* root structure is characterized by a taproot or several main roots that can reach 3 - 4.5 m and these can form a large quantity of biopores, increase soil organic carbon and improve water flow in the soil[13,14].

Strategies such as habitat restoration have been increasingly implemented to conserve and reintroduce native species like *Silphium;* however, little is still known about the genetic diversity of the genus, with only one phylogenetic study published to date[15]. The lack of genetic studies and genomic resources in wild species with agricultural potential like *Silphium* underscores an urgent need to develop the biological framework required to advance *de novo* domestication. Genetic variation for important traits necessary for the development of a new crop may be available in germplasm collections; however, to effectively use these resources in breeding programs, it is necessary to apply novel approaches that integrate genomic information with the phenotype[8,16]. Genomic studies on the diversity of available germplasm and the use of methods such as genome wide association studies and genomic selection will increase the efficiency of a breeding program and may generate improved varieties in a short period of time.

Here we develop and begin to test a framework for rapid *de novo* domestication using *Silphium*: initial genome assembly and annotation enabled the development of low-coverage short-read and target-capture sequencing of wild germplasm accessions, wild relatives, and breeding materials. Efficient genotyping options enabled preliminary analysis of phylogeny, population structure, environmental adaptation, and trait variation. We integrate these methods to reveal the diversity and genomic architecture of domestication traits and offer the most comprehensive genomic analysis for the genus to date, providing critical genomic resources to support genomics-enabled breeding and restoration efforts.

2 MATERIALS AND METHODS

2.1 Plant material

2.1.1 Genome assembly

For the genome assemblies, two individuals, “Bad Astra” (HAP1 - *Silphium integrifolium*) and “1JHW” (HAP2 – *Silphium perfoliatum*) were selected for sequencing along with their F_1 interspecific hybrid (3ZZP) for a haplotype-phased assembly in a trio binning approach[17] (Supplementary Figure S1 and S2). “Bad Astra” is a genet developed at The Land Institute (TLI) for vigor, large heads with numerous ray florets, and partial disease resistance. “1JHW” is a

clone of a *S. perfoliatum* genet grown at TLI from seed collected from a wild population in Nebraska, USA.

2.1.2 Silphium association panel (SAP)

The *Silphium* association panel (SAP) was developed at TLI by selecting 258 individuals from a larger long-term ex situ collection. The genets were grown from seed collected in wild *Silphium* populations in several states of the United States (Supplementary Figure S1, Supplementary File S1). The SAP is composed of 225 *S. integrifolium*, 25 *S. radula*, 3 *S. perfoliatum* genets and 5 *S. integrifolium* × *perfoliatum* hybrid genets. Mature genets were cloned by crown division and replicate clones were planted in the HudsonAlpha experimental field in Huntsville, AL, USA and at TLI in Salina, KS, USA). The experimental design was a lattice design with two complete replications and six incomplete blocks per replication. Twelve individuals had additional replications (5 to 12) randomized across the fields to serve as additional checks, resulting in 600 plots in each location. An initial randomization was performed within the replication and a second randomization was performed within each block. After transplanting in the field, 85 of the 600 individuals died in Alabama (14%) and replacements for these genets were planted two weeks later. To account for potential effects of this later transplant, the transplantation date was included as covariate for the analysis of all traits evaluated in Alabama 2024. Phenotyping was performed in the field in the four environments (AL24, AL25, KS24, KS25) (see section 2.12).

2.1.3 Silphium phylogeny

To expand our knowledge about the evolution and species relationships in *Silphium*, 41 samples from 14 different *Silphium* species, and eight samples from three closely related members of the Asteraceae family as outgroups (*Lindheimera texana, Engelmannia peristenia and Helianthus maximiliani*) were obtained from collaborators or commercial nurseries (Supplementary Figure S1, Supplementary File S2).

2.2 DNA and RNA extraction

Young leaves from the F1 hybrid 3ZZP were collected and flash-frozen in liquid nitrogen for high molecular weight (HMW) DNA extraction for PacBio long read sequencing and nuclei isolation for Dovetail Omni-C library preparation. HMW DNA isolation was performed in the Arizona Genomics Institute. One gram of frozen young leaf tissue was used as input for a Dovetail Omni-C library preparation following the manufacturer's protocol. Leaf tissue from the parents, Bad Astra and 1JHW, was collected and DNA was extracted using the modified CTAB extraction [18] for Illumina short read sequencing. For the *Silphium* Association Panel and the

phylogenetic studies, DNA was extracted from lyophilized leaf tissue using the EchoLUTION Plant DNA Kit (BioEcho Cologne, Germany).

RNA was extracted from the F1 hybrid 3ZZP to generate RNAseq data for gene annotation when this plant had several stalks and was actively flowering with multiple inflorescences at different stages of development (Supplementary Table S1). Eleven tissue types were sequenced: From capitula, stage 5: disk floret pales (bracts), disk florets un-swollen, disk floret pre-anthesis, pre-anthesis ovaries, phyllaries (involucral bracts); From heads, stage 7: mature unpollinated ovaries; From fully expanded leaves: lamina, mid-vein; From the root system: thick brown woody roots; From main stalks, green and healthy: stem nodes and internodes. Tissues were ground using a Freezer Mill, and the RNA was extracted with CTAB [19]. The Zymo Clean and Concentrator kit was used for RNA cleanup after extraction.

2.3 Genome sequencing and Assembly

CCS reads (57X coverage) were generated with the PacBio Sequel II instrument for the hybrid 3ZZP. Sequencing of the Omni-C libraries for 3ZZP (47X coverage) and the Illumina TruSeq DNA PCR-free libraries for the two parental lines Bad Astra and 1JHW (50 X coverage) was performed on an Illumina NovaSeq6000. All sequencing was performed at the HudsonAlpha Institute for Biotechnology.

Initial estimation of the nuclear genome size was performed with jellyfish (v2.3.10)[20] to count k-mers and visualization with GenomeScope2.0 and Smudgeplot (v0.4)[21] (Supplementary Figure S3). The short-read sequences from the parental genotypes were used to phase the two haplotypes using HiFiasm (v0.19.8)[22] following the trio binning method[17]. Dovetail Omni-C reads were separately mapped to the HAP1 and HAP2 contig sets using bwa-mem[23] and chromosome scale scaffolding was performed with Juicer[24]. The contigs were then oriented, ordered, and joined together into chromosomes per haplotype using the Omni-C data. Each chromosome join was padded with 10,000 ‘N’ calls. The assemblies were subsequently polished using RACON[25].

Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the production assembly. The remaining scaffolds were screened against bacterial proteins, organelle sequences, and GenBank nr, and removed if found to be a contaminant. A BUSCO (v5.8.2) analysis was performed to assess the completeness of the genome, using the *eudicots_odb10* database[26]. The genome structure and synteny were visualized using the R package Genespace[27].

2.4 Gene annotation

Illumina TruSeq stranded RNAseq libraries constructed for the F1 Hybrid 3ZZP were sequenced at the HudsonAlpha Genomic Sequencing Center with an Illumina NovaSeq6000. The RNA sequencing reads from the eleven different tissue types described above (section 2.2) were used for gene prediction in BRAKER3[28]. The RNAseq data was used in conjunction with protein sequences from four species in the Asterid clade: Carrots *Daucus carota* (Dcarota_388), Lettuce *Lactuca sativa* (Lsativa_467), Sunflower *Helianthus annuus* (Hannuus_494) and Tomatoes *Solanum lycopersicum* (Slycopersicum_796). The protein data was downloaded from *https://phytozome-next.jgi.doe.gov/.* BRAKER3 was run with the following parameters: "--threads 8 --softmasking". After completion, the resulting *gtf* file was further filtered with the auxiliary python script from BRAKER3 *selectSupportedSubsets.py* to select genes fully supported by the RNAseq and the protein database. Genes were further inspected with the script *generateReport.py* to visualize the distribution of gene lengths. The functional annotation of the genes was performed with InterProScan version 5.66-98.0[29].

2.5 Chloroplast assembly and annotation

The chloroplasts from *S. integrifolium* and *S. perfoliatum* were assembled with the OatK pipeline (https://github.com/c-zhou/oatk) utilizing the angiosperm database, which is the closest database available in OatK. The chloroplast genome was then polished using RACON[25]. The assembled genomes were annotated using GeSeq v2.0.3[30] and visualized using OGDRAW v1.3.1[31], both tools available at https://chlorobox.mpimp-golm.mpg.de/geseq.html.

2.6 Transposable elements and repeats annotation

Transposable elements prediction and annotation was performed with the Extensive de novo TE Annotator (EDTA) pipeline (v1.9.3)[32] using the following parameters: "--overwrite 0 --sensitive 0 --anno 1 --evaluate 0 --threads 25".

2.7 Comparative genomics

We compared the gene families in *S. integrifolium* and *S. perfoliatum* with sunflower (*Helianthus annuus*) and *Scaevola taccada*, a member of Goodeniaceae that is one of the closest outgroups of Asteraceae family. We retrieved the genome assemblies of two sunflower genomes, the XRQ.v2 and HA412-HO.v2 assemblies[33,34] and the *Scaevola taccata* assembly ST1.0[35]. To ensure consistency among the different datasets, we annotated the repeats and the genes for the genomes, using the same methods used for Silphium with EDTA, BRAKER3 and InterProScan. For the gene annotation, RNAseq data from 8 different tissues (leaves, roots, stem, bract, ovary,

stamen, style and ligule was retrieved for sunflower XRQ (SRP092742) and for *Scaevola taccata*, RNAseq data from 4 different tissues (flower, leaves, stem and root) were used for the gene annotation.

2.8 SAP sequencing and variant calling

The *Silphium* association panel of 258 accessions was sequenced with two combined strategies. The first was low-coverage sequencing targeting an average depth of 1X. Libraries were prepared with the Twist Bioscience 96-Plex kit and sequencing was performed on the Illumina NovaSeq 6000. The second strategy was target-sequencing aiming to enrich sequencing in genic regions containing SNPs. To identify the SNPs, reads from the low-coverage sequencing were mapped to the "Bad Astra" genome with bwa mem[23]. SNP calling and filtering were performed with Bcftools[36] with filtering parameters: MAF > 0.01, MAPQ > 10, depth ≥ 5, missing < 95%. The missing threshold was purposely relaxed to allow the maximum number of SNPs existing in the population to be sampled for target sequencing. A total of 77 million SNPs were identified and were then annotated with SnpEff to allow the classification of variants according to their predicted impact in the gene function[37]. The SNPs were sequentially selected for probe design based on their predicted effect, where 13,977 SNPs with high impact, 12,968 with moderate impact and 745 with low impact were chosen. Another 2,594 SNPs classified as modifiers were also included. An additional 15,101 SNPs identified outside genes but up to 5,000 bp from the transcription start site were targeted. To design probes, the "Bad Astra" genome was used as a reference and a bed file with the SNP coordinates was produced. A custom script was developed to extract a sequence of 300 bp centered on the target SNP and design 120 bp probes. The script applies filters for BLAST matches to ensure uniqueness and estimate GC, annealing temperature, hairpin and homodimer formation (https://github.com/renanso/enrich/). In total, 50,000 probes were designed, then synthesized by Twist Bioscience and used in the Twist target enrichment standard hybridization (Supplementary File S3). For the enrichment, sequencing libraries were prepared with Twist Bioscience Flex Prep and after hybridization, the 258 accessions were sequenced on an Illumina Novaseq 6000 (Supplementary Figure S4).

The two datasets, low-coverage and target sequencing, were combined for each individual and then processed for variant calling. The calling was based on the Khufu platform to generate a hapmap[38]. Khufu is a pipeline that integrates read mapping with bwa[23], variant calling and quality control with bcftools[36] and imputation with Stitch[39]. SNPs were filtered with the following parameters: number of alleles =2, sequencing depth ≥ 3 X, minor allele frequency

(MAF) ≥ 0.01, missing calls per loci < 30%. In the end, the hapmap retained 449,554 SNPs (Supplementary Figure S5).

2.9 Reconstructing the phylogeny of *Silphium*

To reconstruct the genus phylogeny, the 49 samples sequenced using the target sequencing method were analyzed together with the two sunflower genomes HA412HO and XRQ, *Scaevola taccada*. In addition, six *S. radula,* three *S. perfoliatum,* six *S. integrifolium* and five *S. integrifolium* × *perfoliatum* from the SAP were included, resulting in 72 samples in total for this analysis.

To select the most informative loci from the 50,000-target sequencing panel, the top 7,000 loci with reasonable depth (15X < depth < 150X) were selected and a sequence of 500 bp centered on the target SNP was extracted using the function getfasta from bedtools 2.28.0[40]. To assemble the target region across the samples, the fastq data was mapped to the targets fasta sequences using HybPiper[41]. To identify paralogs, the command HybPiper *paralog_retriever* was used, and the paralogs were removed from the panel. In addition, only the genes that were assembled in at least 58 of the 72 samples (80%) were selected for alignment. After these steps, the final panel contained 789 loci (Supplementary File S4). To align the sequences, MAFFT[42] was used with the following parameters: --preservecase --maxiterate 1000. Trimal[43] was subsequently used to remove poorly aligned sections from the aligned sequences with -gt .5. To build the gene trees, IQTree version 2.3.6[44] was used with the ModelFinder Plus option[45] and 1,000 bootstraps. To collapse the gene tree with branches below 50% bootstrap value, the nw_ed tool from Newick Utilities[46] was used with the parameter b<50. To build the species tree from the gene tree, ASTRAL[47] was used with no options. Tree files were visualized using FigTree v1.4.4[48].

2.10 Kinship, population structure and ancestry

To analyze the kinship and population structure among the accessions, the kinship matrix based on the centered identity by state and the principal component analysis (PCA) was calculated for all 258 individuals based on all 449,554 SNPs using TASSEL[49]. PCA results were plotted with the R package ggplot2[50] and the hierarchical cluster based on the kinship matrix was plotted with the R package ggtree[51]. Then we conducted a spatial ancestry analysis for *S. integrifolium* using the SNP data set and the GPS coordinates of the sites of collection of 156 individuals. The analysis was performed with the R package tess3r[52], which enables the estimation of ancestry proportions and allele frequencies of the accessions, representing the

changes geographically. Ancestry coefficients were plotted with a fit of interpolating surfaces calculated with a theta value=10 in the Fields Krig Model.

2.11 Genome-environment association (GEA) study

To identify genomic regions associated with environmental variables, we performed a genome scan on two environmental indexes, heat index (HI) and aridity index (AI). For HI, mean temperatures were obtained for the 183 georeferenced genets from Worldclim https://www.worldclim.org/ at a resolution of 30 arc-seconds (1 km^2) considering the historical averages from 1970 to 2000 (Supplementary Figure S6). Data extraction and preparation were performed using the R packages geodata, terra and raster [53–55]. The HI was calculated using the evapotranspiration model[56,57], considering the mean temperature for the months of June, July and August, which are the months of the reproductive phase of *Silphium* and when plants tend to be more sensitive to high temperatures (Equation 1).

$$HI = \sum_{i=1}^{k} \frac{Tm^{1.514}}{5} \qquad (1)$$

Tm is the 30-year average of the mean monthly temperature and k is the number of months.

The AI values were retrieved from the Global Aridity Index and Potential Evapotranspiration Database (Global-AI_PET_v3)[58] for each of the georeferenced individuals. The AI corresponds to the ratio between precipitation and potential evapotranspiration, and it measures moisture availability for plant growth. The Global AI database provides data at a 30 arc-seconds (1 km^2) resolution averaged monthly for the period of 1970–2000. The average AI values for the summer months of June, July and August was calculated and the GEA analysis was performed with the 183 individuals that had GPS coordinates available and passed the sequencing quality control.

The association between the indexes and the genotypes was performed using the R package GAPIT3 [59,60], using two different statistical models (FARMCPU and BLINK). The assumption for the GEA analysis is that an organism that occupies and thrives in a specific environment has the necessary adaptations to the climatic conditions of that environment (e.g. higher mean temperature, lower precipitation) and these adaptations are the result of mutations that can be captured as signals in association studies.

2.12 Genome-wide association study

A genome-wide association study (GWAS) analysis was performed with the 449554 SNPs for four traits: Ray floret count (RFC), seed mass (SM), seed number per capitulum (SNC) and receptacle diameter (RD). RFC was determined by counting the ray florets in the first

capitulum at complete anthesis. For SM, the mass of 20 seeds was measured and the average mass per seed in milligrams was determined. For SNC, the total number of seeds was counted in the most representative mature capitulum. RD was measured with a caliper in millimeters (Supplementary Table S2). Best linear unbiased estimation (BLUE) was calculated in single and the combination of environments (AL24, KS24, AL25, KS25, Combined) using the R package lme4 [61] with the following model:

$$y_{ijkl} = \mu + g_i + b_{j(k)} + r_k + e_l + (ge)_{il} + \varepsilon_{ijkl} \qquad (2)$$

y_{ijkl} corresponds to the response value for the i^{th} genotype in the j^{th} block within k^{th} replication within l^{th} environment. μ, g_i,$b_{j(k)}$,r_k,e_l and ge_{il} represent the effect of the mean, the genotype (fixed), the incomplete block (random), replicate (random), the environment (random) and the interaction genotype × environment (random), respectively. ε_{ijkl} is the intra-block residual, assumed to be normally and independently distributed with mean 0 and variance σ^2. BLUEs calculated for the individual environments followed the model above without the effect of the environment e_l and the interaction $(ge)_{il}$.

The variance components were determined in a model considering all factors as random effects using the restricted maximum likelihood method (REML). These components were used to determine entry-mean based heritability estimates according to the following equation [62]:

$$H^2 = \frac{\sigma_g^2}{\left(\sigma_g^2 + \frac{\sigma_{ge}^2}{e} + \frac{\sigma_\varepsilon^2}{r \times e}\right)} \qquad (3)$$

H^2 is the broad-sense heritability, σ_g^2 is the genotypic variance, σ_{ge}^2 is the G×E variance, σ_ε^2 residual variance, e is number of environments and r is the number of replicates.

The same models used for the GEA (FARMCPU and BLINK) were implemented in the GWAS. The analysis accounted for the population structure captured by the first three principal components. All analyses conducted with R packages were performed in the R version 4.4.2 [63]. The identified loci were classified as genic, 10-Kb window, 50-Kb window or intergenic (> 50 Kb from a gene), depending on where the most significant SNP was located. The loci identified were further validated with a t-test comparing the effect of the alleles at each locus on the phenotype. To understand the effect of the locus on the phenotype, the SNP genotypes at each identified locus were retrieved and tested as predictors for the phenotype in a linear regression.

3 RESULTS

3.1 Assembly, annotation and structure of the *S. integrifolium* and *S. perfoliatum* genomes

The genome assembly resulted in 332 contigs for *Silphium integrifolium* 'Bad Astra' and 241 contigs for *S. perfoliatum* '1JHW' and the scaffolding resulted in 7 chromosomes per haplotype. Contig N50 size was 41.3 Mb for Bad Astra and 54.7 Mb for 1JHW. The assembly resulted in a genome size of 7.6 Gb for Bad Astra and 7.5 Gb for 1JHW. The BUSCO assessment indicated a high-quality assembly with a completeness score of 97% for *S. integrifolium* (Bad Astra) and 98% for *S. perfoliatum* (1JHW) (Table 1). The LTR assembly index (LAI) was 14.46 for *S. integrifolium* and 14.86 for *S. perfoliatum*. A LAI score of 15.5 has been indicated as a high-quality assembly[64].

The genomes of *S. integrifolium* and *S. perfoliatum* are both composed of seven pairs of chromosomes[65], despite having a large genome. *Silphium* chromosomes are extremely long, with the length of chromosome 1 reaching 1.81 Gb in *S. integrifolium* and 1.79 Gb in *S. perfoliatum* (Supplementary File S5). The genomes of both species are highly heterozygous, with *S. integrifolium* and *S. perfoliatum* having 2.24 and 1.44 % heterozygous loci, respectively (Supplementary Figure S3). The chloroplast genome assembly resulted in a sequence of 152,169 bp for both genomes with 113 genes annotated (Supplementary Figure S7).

When analyzing the Omni-C contact map, we observed the usual main diagonal indicating that each point along the primary DNA structure (sequence) of a chromosome makes the most contacts with the points flanking it (Figure 1A, 1B). Unexpectedly, we observed a weaker pair of diagonals in parallel with the main diagonal (Figure 1E) each offset approximately 43 Mb from the main, central diagonal (Figure 1C, 1D). Another pair of even weaker diagonals was discernible, each tertiary diagonal approximately 43 Mb outside of the secondary diagonal (Figure 1C, 1D). These additional diagonals indicate that each point in the primary sequence often contacts points 43 Mb on either side of it (and, less frequently, 43 Mb from those points). This pattern can be explained by hypothesizing that the DNA has a tertiary structure: a helix with 43 Mb per turn (Figure 1F) that is maintained in both *Silphium integrifolium* and *Silphium perfoliatum* genomes during interphase.

We analyzed the patterns of synteny among the two *Silphium* genomes, two *Helianthus annuus* genomes and *Scaevola taccada*. Overall, there was a strong collinearity between the *S. integrifolium* and *S. perfoliatum*. There were two major inversions, one on chromosome 1 spanning 43 Mb (1055 Mb to 1098 Mb) and another on chromosome 5 spanning 50 Mb (747

Mb to 797 Mb) (Figure 2A). When comparing between the *Silphium* and the sunflower genomes, it was observed that the genomes have low collinearity, where one *Silphium* chromosome splits into several matching sunflower chromosomes. *Silphium* chromosome 1 splits into four pieces that match sunflower chromosomes 3, 10, 12 and 17 and the sunflower chromosome 14 matches entirely to one arm of chromosome 3 in *Silphium* (Figure 2B).

The transposable element (TE) content ranged from 73.5% of the genome in *Scaevola taccada* to 84.97 % in *H. annuus* HA412HO (Figure 2C-E). The content in *S. integrifolium* was estimated at 76.1% and in *S. perfoliatum* it was estimated at 75.7%. LTR elements predominated in all genomes, ranging from 57.9% in *Scaevola taccada,* to 74% in *H. annuus* HA412HO. In *S. integrifolium*, LTR retrotransposons corresponded to 64.1% of the genome and the transposons in the Ty3/Gypsy family were the most common (27.4%). In *S. perfoliatum*, LTR accounted for 63.9 % of the genome and the Ty3/Gypsy family was also the most predominant, corresponding to 27.2%. The genome of *S. integrifolium* is 134 Mb bigger than *S. perfoliatum* and 126 Mb of this difference is due to repetitive elements (94%) (Table 2).

We evaluated whether the differences in genome sizes between *Silphium*, sunflower and *Scaevola* would also correspond to differences in gene number. There was not a clear trend of having a larger genome resulting in more genes, with *H. annuus* XRQ having the highest count (49,286), followed by *H. annuus* HA412HO (45,087), *S. integrifolium* (44,653), *S. perfoliatum* (43,390) and *Scaevola taccada* (18,435) (Figure 2F). For the functional annotation of the genes with InterProScan, *Scaevola taccada* had 99.88% of the genes assigned to known protein families, followed by *H. annuus* HA412HO (79.61%), *S. perfoliatum* (78.85%), *S. integrifolium* (78.10%), and *H. annuus* XRQ (77.31%) (Supplementary File S6).

Silphium integrifolium had the highest count for genes in the cytokinin pathway (16), and *S. perfoliatum* had the highest count for genes associated with histones (236) and squalene pathway (22) (Figure 2G). *Helianthus annuus* XRQ had the highest count for leucine rich repeat genes (154), Self-incompatibility genes (27) and Auxin (203). *Helianthus annuus* HA412HO had the highest count for genes involved in fatty acid metabolism (55) and in the ethylene pathway (54). *Scaevola taccada* had the lowest number of genes in all the pathways investigated.

3.2 *Silphium* phylogeny

We reconstructed the phylogeny of *Silphium* aiming to resolve the relationship between 14 species using a set of 789 loci (Figure 3A). *Silphium* and sampled taxa in subtribe Engelmanniinae were both recovered as monophyletic with complete support. As expected, *Silphium* species formed two clades variously treated as subgenera or sections: *Composita* and

Silphium. *Composita* includes species with large taproots and persistent basal leaves: *S. albiflorum* Gray, *S. laciniatum* L., *S. pinnatifidum* Elliott and *S. terebinthinaceum* Jacq. In this clade, *S. pinnatifidum* and *S. terebinthinaceum* formed an intermixed group sister to an assemblage including *S. albiflorum* and *S. laciniatum*. Subgenus *Silphium* contained *S. asteriscus* L., *S. connatum* L., *S. integrifolium* Michx., *S. mohrii* Small, *S. perfoliatum* L., *S. perplexum* J.R.Allison, *S. radula* Nutt., *S. simpsonii* Greene, *S. trifoliatum* L. and *S. wasiotense* Medley.

Within subgenus *Silphium*, *S. integrifolium* formed an intermixed group with *S. simpsonii, S. radula, S. asteriscus, S. trifoliatum* and *S. mohrii*. Also in this section, *S. connatum*, *S. perfoliatum*, and a single sample identified as *S. perplexum* clustered together (Figure 2C). *Silphium wasiotense* was supported as a sister of this clade but with low support (0.43). Lastly, *S. mohrii* was weakly supported as sister of all other species in subgenus *Silphium* (bootstrap support = 0.45; Supplementary Figure S8).

While we managed to recover a strong backbone of the genus including the two subgenera, support values within subgenus *Silphium* were quite low; only a sister relationship of *S. connatum* and some *S. perfoliatum* was robustly supported (support = 0.98; Supplementary Figure S8). More interestingly, numerous species across both subgenera were recovered as polyphyletic, with *S. terebinthinaceum* and *S. pinnatifidum* both nonmonophyletic within an intermixed clade and *S. integrifolium* and *S. radula* appearing in various positions throughout subgenus *Silphium*. *Silphium asteriscus* was also polyphyletic with regard to *S. trifoliatum*, although *S. trifoliatum* is often considered a variety or subspecies of *S. asteriscus* as a result of ongoing taxonomic debate[7].

3.3 Genetic diversity and population structure in the *Silphium* association panel

The 449,554 SNPs identified in the SAP showed a uniform distribution across all chromosomes in *Silphium integrifolium* and revealed a pattern of fast linkage disequilibrium decay in this population, with most linkage disappearing after 1 kb (Supplementary Figure S5). The hierarchical clustering based on the kinship of all 258 genets in the SAP indicated a top-level division of the population in two major groups (Figure 3B). The PCA plot further corroborates the hypothesis of two populations when the points were color coded based on longitude, when it became evident that the genetic relatedness is strongly correlated with the population location in the east-west gradient (Figure 3C). The spatial ancestry analysis for 156 *Silphium integrifolium* genets with GPS coordinates was conducted varying the number of ancestral populations from 2 to 5. With K=2, there is a clear east-west division that persists even

when more ancestral populations are considered (Figure 3E). When analyzing the individual ancestry coefficients, we observed several individuals with mixed ancestry, indicating possible events of hybridization between individuals from east and west (Supplementary Figure S9). In fact, the SAP exhibited a high level of heterozygosity, with 14.9% of the individuals showing a heterozygous marker per locus on average across the 449,554 SNPs. When the ancestry models use the assumption of three ancestral populations (K=3), a third ancestry group appears in eastern Illinois and Wisconsin. In this context, the SAP can be further divided into four major groups: *S. integrifolium* from the east, *S. integrifolium* from the west, *S. integrifolium* from northeast and *S. radula* (Figure 3D). Taken together these results suggest that clinal variation occurs mostly in the East-West gradient, with two major populations existing at the present time. Interestingly, the *S. radula* genets grouped together with the *S. integrifolium* from the east population, based on the kinship calculated by identity by state.

3.4 Phenotypic data analysis of the *Silphium* diversity panel

The four phenotypic traits, ray floret count (RFC), seed mass (SM), seed number per capitulum (SNC) and receptacle diameter (RD) that were evaluated in both locations (AL and KS) showed an overall high heritability, ranging from 0.74 to 0.92 (Supplementary Table S3). A BLUE based correlation analysis was performed to determine possible associations between environmental variables and phenotypic traits. All four capitulum traits (RFC, SM, SNC, and RD) showed a moderate to strong correlation among themselves (0.35 to 0.81) and a negative correlation to the aridity index, with values ranging from -0.28 to -0.58 (Figure 4A, Supplementary Figure S10). The aridity index value is low in drier and high in wetter environments. Based on this, it is possible to infer that that genotypes adapted to drier locations tend to produce heavier and more numerous seeds per capitulum in a bigger receptacle.

3.5 Genome-environment and Genome-wide association study

The genome scans for association with environmental indices and phenotypic traits identified 81 loci across all chromosomes, each explaining from 1% to 33% of the phenotypic variance. There were 63 loci explaining more than 5% of the phenotypic variance and 44 of these loci explained more than 10% (Supplementary Table S11, Supplementary Figure S13). For the aridity index, two intergenic SNPs explained more than 5% of the variance. Chr02_892024560 was detected in the FarmCPU model with 5.87% and the other SNP Chr06_274305900 was detected in both FarmCPU and BLINK with 12.75%. For heat index, nine loci were identified, with one of these (Chr03_78530134) being detected in both GWAS models, explaining 12.44% of the variance (Figure 3F-I).

For average seed mass, 11 loci were identified, together explaining 56.04% of the variance. The SNP with the strongest association with average seed mass was Chr01_1030878591 with 9.03% of the variance and characterized by a SNP in the gene silint01g09207 (Pinoresinol reductase-related protein). For seed number per capitulum, 25 loci were identified, ranging from 1.94% to 25% of the variance. Notably, the SNP Chr01_1583902655 in the gene silint01g14554 (MATE family efflux transporter) was detected with both FarmCPU and BLINK in the KS25, accounting for the largest percentage of the variance (25%), and being also associated with ray floret count and receptacle diameter (Figure 4B-D).

For ray floret count, 28 loci were detected explaining from 1.39% to 33.59% of the variance. Similar to seed number per capitulum, the SNP Chr01_1583902655 in the gene silint01g14554 was detected by BLINK and FarmCPU both in the combined analysis and in KS24, explaining 29.39% of the variance. Another two SNPs (Chr01_695173749 and Chr04_222292225) were also detected in both GWAS models, explaining 22.82% and 23.03% of the variance. Lastly, the SNP Chr03_721440854 was detected for the trait in AL25 with an associated variance of 31.19% (Figure 4B-D). For receptacle diameter, 11 loci were detected, ranging from 2.65% to 30.52% of the variance. Two SNPs that were previously identified for seed number per capitulum and ray floret count, also explained a large portion of the variance for this trait, with Chr01_1583902655 and Chr03_721440854 explaining 24.29% and 28.47%, respectively (Figure 4B-D). Since the three capitulum traits (seed number per capitulum, ray floret count and receptacle diameter) formed a group of high correlation and the GWAS results showed that they had four loci in common detected in multiple models and environments (Chr01_695173749, Chr01_1583902655, Chr03_721440854, Chr04_222292225) we decided to further investigate these loci to understand how they affect inflorescence morphology.

When grouping the genets in the SAP into wild collections and breeding material and comparing the alleles at each SNP site, we observed that in the wild collections, the favorable alleles exist majorly in the heterozygous state for the four loci, with only one occurrence of fixed favorable allele in wild accessions (AA genotype in SNP Chr03_721440854 in the accession S005T6) (Figure 4E). Interestingly, there are three *Silphium radula* individuals, S005QM, S005TE collected in Oklahoma (heterozygous in SNP Chr01_1583902655) and S005R9 from Texas (heterozygous in SNP Chr03_721440854, which shows that the mutations are not exclusive to *Silphium integrifolium*. The four loci start to appear in the homozygous state for the favorable alleles only in the breeding genets (Figure 4E). This is also evident by comparing the floret count between wild and breeding genets (Figure 4F). The higher frequency of individuals

with multiple favorable alleles in the breeding group is likely a result of crossing and selection for high floret count, and consequently higher seed number per capitulum. In the wild collections subgroup of the SAP, there are eight allele combinations in the four loci (haplotypes): the baseline wild type (GTGG) and seven mutant haplotypes (GHGG, GTAG, GTGH, GTHG, GTHH, HTGG and HTHG). Most of the mutant haplotypes are present in several of the natural populations sampled, with the GTHG haplotype being the most widespread with occurrences in Nebraska, Oklahoma, Texas and Kansas. GHGG occurs in Oklahoma, Illinois and Arkansas, GTAG in Nebraska and Mississippi, HTGG in Nebraska and Kansas and HTHG occurs in Kansas and Mississippi. The haplotypes GTHH and GTGH occur only in Kansas (Figure 4G). The combined effect of the favorable alleles at the four loci is evident when comparing the images of capitulum from individuals with wild type alleles (GTGG) versus those with different mutations (Figure 4H).

4 DISCUSSION

In this study, we report the first sequenced genomes of two *Silphium* species and employed this new resource to genotype different populations for phylogenetic studies, population structure and mapping loci associated with domestication. A phased reference genome was assembled based on an interspecific cross between *S. integrifolium and S. perfoliatum.* The genome assembly confirmed that *Silphium* has seven chromosomes, as previously indicated by cytological studies[65,66]. The genomes are large: 7.6 Gb for *S. integrifolium* and 7.5 Gb *S. perfoliatum,* of which approximately 75% are transposable elements.

When inspecting the assemblies, we observed that a second diagonal and traces of a third one were present in the contact matrix. The same phenomenon has been observed recently in *Paris polyphylla*, a member of the Melanthiaceae family with genome size of 54.58 Gb/1C and the largest plant chromosome assembled to date, reaching 14.14 Gb[67]. The authors argued that the secondary diagonal is evidence for a helical structure unique to large chromosomes in interphase nuclei, which exists to prevent DNA entanglement and promote efficiency for transcriptional regulation. This seems that it is a result of formation of consecutive loops emanating from a central axial scaffold of the chromosome independently of cell type or stage. These observations need to be validated in additional analysis of different cell stages to elucidate the formation of this helical structure in large chromosomes. It has been also observed that several genes involved in DNA repair and chromatin structure maintenance such as histones were enriched in *P. polyphilla*[67]. Here we observed that *S. perfoliatum* had the highest number

of histone genes (236), although the difference was not substantial in comparison to the sunflower assemblies (HA412HO = 224, XRQ = 226).

Aiming to examine relationships between *Silphium* species, we reconstructed the phylogeny with 14 species, including multiple samples for most taxa. Overall, the phylogeny had great accordance to the first *Silphium* phylogeny based on ITS and ETS sequencing data[15], with a clear division into 2 sections or subgenera, *Composita* and *Silphium*. Subgenus *Composita* had similar structure, with *S. albiflorum* and *S. laciniatum* forming a clade and *S. terebinthinaceum* and *S. pinnatifidum* intermixed together.

A major difference between the previous phylogeny and our study was a lack of resolution in the *Silphium* section. *S. integrifolium*, *S. wasiotense* and *S. perfoliatum* were previously grouped together and *S. radula, S. asteriscus* and *S. trifoliatum* formed another group[15]. Our analysis indicates that *S. integrifolium* is more closely related to this latter group, instead of *S. wasiotense* and *S. perfoliatum.* However, relationships between species across our tree were poorly supported, with several species appearing as polyphyletic, especially *S. integrifolium*, *S. radula*, *S. terebinthinaceum*, and *S. pinnatifidum*. *Silphium* species exhibit high morphological and ecological variability across large geographic ranges, making their taxonomy difficult to ascertain[7]. This idea is mirrored in our phylogenetic findings, as the polyphyletic species are among the most widespread and most variably treated in the taxonomic literature in terms of varieties, subspecies, or recognition of entirely separate species [7]. Further development of taxonomic and evolutionary knowledge of *Silphium* must be prioritized in order to indicate possible candidates for the exploration of additional gene pools that could be accessed as the domestication of *Silphium* species advances and to inform conservation strategies for wild *Silphium* populations. For this, the curated set of 789 loci developed here is an important tool to capture variation across the genus.

From the 258 genets in the SAP, 183 were collected in natural populations and from these, 157 were *S. integrifolium*. Using the GPS coordinates and genetic information from the 449,554 SNPs, we tested spatial ancestry algorithms to analyze the ancestry of *S. integrifolium* populations and observed that the populations are derived from two major groups split along the east and west gradient. When increasing the number of possible populations in the model to three (K=3), a group of individuals form a distinct population in the northeast. This population coincides with the endemic occurrence of *Silphium integrifolium var. neglectum* across Illinois, Wisconsin and Indiana, indicating a possible differentiation of this variety from the other populations[68]. When assuming K=4, one individual in Texas appears to represent a distinct southwest population, which approximates with the three groups identified previously[69], one

eastern group (Arkansas, Illinois, Mississippi, Missouri, and Wisconsin), one southern (Oklahoma and Texas), and a western (Kansas and Nebraska). A considerable admixture in individuals from the western and southern populations has been observed[69], which would explain the merge of these two populations when K=2. The confirmation of these sub-specific lineages will make future efforts to diversify *S. integrifolium* collections, genetic panels, and breeding populations more efficient. The identification of admixed populations may represent an opportunity for breeders or conservationists to discover novel, adaptive haplotypes[70], potentially accelerating breeding for traits such as disease resistance.

The genome assembly enabled the large-scale genotyping of the diversity panel, and the identification of significant regions associated with the aridity and heat index, seed mass, seed number per capitulum, ray floret number and receptacle size. GWAS performance is greatly improved when the reference genome used is a complete genome assembly instead of a draft[71]. A complete assembly also allows the correct identification of candidate genes when performing genome scans. This is the first ever genome-wide study reported for the genus *Silphium*, revealing with great detail the genetic architecture of four capitulum traits and two environmental indexes. The loci Chr06_274305900 and Chr03_78530134 were identified in both GWAS models and explained 12.75% and 12.45% of the variance for AI and HI, respectively. Although these loci are found in the intergenic space and there is not a clear explanation on how they correlate with the environmental indexes, they can be useful as targets for selection to ensure that *Silphium* continues to exhibit adaptability to stress, as domestication advances.

Four major loci associated with ray floret count, number of seeds per capitulum and receptacle diameter were detected in this study. The ray floret count is particularly important because the capitulum of *Silphium* has a central disk with the staminate florets (male) surrounded by an outer ring of pistillate ray florets (female) (Supplementary Figure 11). Unlike sunflowers, in *Silphium* only the ray florets are female-fertile and produce seeds, while the disk florets produce only pollen [72]. Therefore, one of the major goals in *Silphium* domestication is to "feminize" the capitulum and increase the number of female florets. In this way, selection for plants with greater numbers of ray florets can potentially increase seed yields [73]. Targeting the four loci identified here for selection, should help accomplish this goal. Most importantly, because of positive correlations, selecting favorable alleles for ray floret count and seed number per capitulum will not negatively affect adaptability to arid environments.

The loci associated with flower and seed are in four genes: silint01g06242 (Alpha/Beta hydrolase), gene silint01g14554 (MATE family efflux transporter), silint03g07384 (Unknown protein) and silint04g01657 (ACR4-related protein). Alpha/beta hydrolases correspond to a

superfamily of proteins with a core structure containing a conserved catalytic triad (Ser-His-Asp), functioning as signaling receptors in plants. In rice, the alpha/beta hydrolase gene *STH1* is involved in protecting cell membrane stability during salt stress and it interacts with the floral integrator gene *Heading date1* to modulate the florigen gene *Heading date 3a* and coordinate flowering time[74]. The multidrug and toxic compound extrusion (MATE) transporters are a family of membrane proteins that facilitate the transport of secondary metabolites, hormones, and xenobiotics to regulate growth, defense, and nutrient uptake. The Arabidopsis MATE family member, DTX50, has been identified as an abscisic acid (ABA) transporter, being involved in ABA mediated growth inhibition[75]. The ACR4 gene is an ACT domain repeat that functions in the control of amino acid metabolism, solute transport, and signal transduction and has shown high expression levels in *Arabidopsis* flowers. ACR gene expression is also affected by different treatments of plant hormones and abiotic stress, but their function is mostly unknown[76].

Despite the knowledge of some functions of the candidate genes identified here, their potential role and the effect of the variants on the capitulum morphology is not clear yet. Therefore, further functional studies are needed to expand our knowledge and validate these candidates. However, even without knowing their role, the SNPs identified can be used in marker assisted selection.

In an attempt to domesticate a new crop rapidly, one might impose a high selection pressure to obtain the top candidates that will suit the best growing conditions. However, this can create a severe bottleneck and consequently causes a loss of genetic diversity and rare alleles, as have been reported in soybean[77]. An important question for those trying to domesticate new crops is: can we obtain significant improvement in agronomic traits without losing genetic diversity? With the resources now available, crop domestication can take a new genomic informed path where it is possible to balance the selection for domestication traits and important adaptation traits from wild relatives, such as tolerance to aridity and heat. This is possible because haplotypes that preserve both domestication alleles (such as the candidate *Silphium* ray floret loci described above) and abiotic stress related alleles (such as candidate *Silphium* aridity and heat adaptation loci) can now be identified at great resolution via sequencing from the start, increasing our ability to track and retain these loci throughout the breeding pipeline.

Even with modern genomic resources, plant domestication remains challenging. However, as seen here with *Silphium*, this framework based on genome assembly, annotation and cost-effective sequencing can enable the use of marker assisted selection and considerably reduce the time to domesticate a new crop. The strategies described here can be seamlessly applied to other wild species, orphan crops with limited resources and major crops with large and complex genomes that need more cost-effective solutions. The assembly of reference

genomes for *Silphium integrifolium* Michx. and *Silphium perfoliatum* L. has been essential for the development of the genotyping strategies used in the populations studied here. These tools were validated by using them to discover environmental and phenotypic associations with markers and candidate genes, and to advance the phylogenetic characterization of the *Silphium* genus. These results give us confidence that both marker-assisted breeding and genomic selection can soon be implemented by *Silphium* breeders. These genomic-assisted breeding methods have proven efficacy in accelerating the domestication of other perennial crops[78].

One question that may arise from the features of *Silphium* genome assembly is: What are the breeding implications for a species with chromosomes revealed to be considerably longer than entire genomes? Recombination in large chromosomes from maize and barley is elevated towards the telomeres and although gene density follows the same pattern in these cases, a considerable number of genes remain within the enormous pericentromeric region where recombination is strongly repressed[79,80]. It should be a priority to survey the recombination landscape for *Silphium* chromosomes for this pattern, given their length, because if present, large, stable pericentromeric haplotypes are likely to have accumulated genetic load like other recombination cold-spots[81] and could have important impact on *Silphium* yield and domestication traits. An affordable strategy to investigate the impact of having larger chromosomes on breeding is to use the markers developed here to characterize the diversity and effects of such haplotypes in existing silphium breeding populations to obtain additional haplotypes with net positive effects. These studies are examples of how early investment in basic genomics and genotyping during new crop domestication makes it possible to discover and maintain types of genetic diversity that could otherwise be difficult to re-introduce at a later stage of domestication.

ACKNOWLEDGEMENTS

The authors thank Patricia Sanmartin, Haley Hale and Zachary Meharg at HudsonAlpha Institute for Biotechnology for assistance in DNA extraction, library preparation, and sequencing. The authors also thank Elijah Nix and Teagan LeVar for assistance in the field phenotyping and Rachel Tavares for helpful feedback during manuscript preparation. Seeds for multiple local *Silphium* species and genotypes were donated by Missouri Botanical Garden and Cooper Breeden at the Southeastern Grasslands Institute.

AUTHOR CONTRIBUTIONS

J.P.C., A.H., Y.B., B.S.H., D.V.T., A.J.M., K.T., and M.J.R. developed the rationale and strategy for building and using Silphium spp. reference genomes. J.S., J.G., L.H.H., J.W. and L.B.B.

generated PacBio, Illumina DNA and RNA and Omini-C data for the genome assembly. J.S., J.J. and C.P. performed genome assembly. R.S. and A.H. conducted genome annotation. R.S, J.P.C. and W.K. generated skim-sequencing and target-sequencing data, performed quality control and variant calling. R.S., D.V.T., J.P.C., B.T., A.B. and E.H. conducted the field trials, and collected phenotypic data. R.S. conducted the comparative genomics, contact mapping, population structure, GEA and GWAS analysis. D.V.T. provided interspecific hybrids and other germplasm. R.S., L.F.N, E.G.M. and D.V.T. selected and obtained plant samples for the phylogeny. L.F.N. supervised the taxonomy of Silphium and interpreted the phylogeny. R.S., A.H., D.V.T. and J.P.C wrote the manuscript draft. A.J.M., B.S.H., Y.B., K.T., L.F.N., M.J.R, E.G.M., S.W. contributed to manuscript preparation and revision. All the authors read and approved the final manuscript.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

FUNDING

Funding for this research was provided by The Land Institute, The Perennial Agriculture Project and The Malone Family Foundation. AH is funded by National Science Foundation IOS-PGRP CAREER award #2239530.

DATA AVAILABILITY STATEMENT

Raw sequencing data and genome assembly have been deposited in the National Center for Biotechnology Information under the project PRJNA1417267. Genome annotations and additional files will be available at Figshare.

5 REFERENCES

1. Ross-Ibarra, J., Morrell, P. L. & Gaut, B. S. Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proc. Natl. Acad. Sci. U. S. A.* **104 Suppl 1**, 8641–8648 (2007).
2. Yu, H. & Li, J. Breeding future crops to feed the world through de novo domestication. *Nat. Commun.* **13**, 1171 (2022).
3. Shelef, O., Weisberg, P. J. & Provenza, F. D. The Value of Native Plants and Local Production in an Era of Global Agriculture. *Front. Plant Sci.* **8**, 2069 (2017).
4. Glover, J. D. *et al.* Agriculture. Increased food and ecosystem security via perennial grains. *Science* **328**, 1638–1639 (2010).
5. Bayer, P. E. *et al.* Sequencing the USDA core soybean collection reveals gene loss during

domestication and breeding. *Plant Genome* **15**, e20109 (2022).
6. Zsögön, A. *et al.* De novo domestication of wild tomato using genome editing. *Nat. Biotechnol.* **36**, 1211–1216 (2018).
7. Brock, M. & Weakley, A. Silphium Linnaeus 1753 (Rosinweed). in *Flora of the Southeastern United States* (ed. Weakley, A. S.) (University of North Carolina Herbarium, North Carolina Botanical Garden, 2020).
8. Van Tassel, D. L. *et al.* Accelerating *Silphium* domestication: An opportunity to develop new crop ideotypes and breeding strategies informed by multiple disciplines. *Crop Sci.* **57**, 1274–1284 (2017).
9. Hwang, H.-S., Liu, S., Winkler-Moser, J., Singh, M. & Van Tassel, D. Composition and oxidative stability of silflower (Silphium integrifolium) seed oil and its potential as a new source of squalene. *Authorea* (2023) doi:10.22541/au.169083365.50710798/v1.
10. Wever, C. *et al.* Towards high-biomass yielding bioenergy crop Silphium perfoliatum L._ phenotypic and genotypic evaluation of five cultivated populations. *Biomass Bioenergy* **124**, 102–113 (2019).
11. Lozano-Grande, M. A., Gorinstein, S., Espitia-Rangel, E., Dávila-Ortiz, G. & Martínez-Ayala, A. L. Plant sources, extraction methods, and uses of squalene. *Int. J. Agron.* **2018**, 1–13 (2018).
12. Nastasi, L. F., Smith, C. N., Tooker, J. F. & Deans, A. R. Catalog of endophytic insects associated with rosinweeds (*Silphium* L., Heliantheae, Asteraceae) in North America. *Psyche (Camb. Mass.)* **2024**, (2024).
13. Weaver, J. E. Classification of root systems of forbs of grassland and a consideration of their significance. *Ecology* **39**, 393–401 (1958).
14. Rohlmann, L., Köhne, J. M., Deiglmayr, K. & Geers-Lucas, M. Perennial Roots, Lasting Structure: How Silphium perfoliatum Alters pore structure to shape carbon storage and water flow. *Geoderma* **463**, 117565 (2025).
15. Clevinger, J. A. & Panero, J. L. Phylogenetic analysis of Silphium and subtribe Engelmanniinae (Asteraceae: Heliantheae) based on ITS and ETS sequence data. *Am. J. Bot.* **87**, 565–572 (2000).
16. Reinert, S. *et al.* Mating compatibility and fertility studies in an herbaceous perennial Aster undergoing de novo domestication to enhance agroecosystems. *Agron. Sustain. Dev.* **40**, 27 (2020).
17. Koren, S. *et al.* De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4277.

18. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *PHYTOCHEMICAL BULLETIN* (1987).
19. Gasic, K., Hernandez, A. & Korban, S. S. RNA extraction from different apple tissues rich in polyphenols and polysaccharides for cDNA library construction. *Plant Mol. Biol. Rep.* **22**, 437–438 (2004).
20. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
21. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).
22. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
23. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
24. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–98 (2016).
25. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
26. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
27. Lovell, J. T. *et al.* GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *Elife* **11**, (2022).
28. Gabriel, L. *et al.* BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. *bioRxiv* (2024) doi:10.1101/2023.06.10.544449.
29. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
30. Tillich, M. *et al.* GeSeq - versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* **45**, W6–W11 (2017).
31. Greiner, S., Lehwark, P. & Bock, R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* **47**, W59–W64 (2019).
32. Ou, S. *et al.* Benchmarking transposable element annotation methods for creation of a

785 streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
786 33. Badouin, H. *et al.* The sunflower genome provides insights into oil metabolism, flowering
787 and Asterid evolution. *Nature* **546**, 148–152 (2017).
788 34. Hübner, S. *et al.* Sunflower pan-genome analysis shows that hybridization altered gene
789 content and disease resistance. *Nat. Plants* **5**, 54–62 (2019).
790 35. Shen, F. *et al.* Comparative genomics reveals a unique nitrogen-carbon balance system in
791 Asteraceae. *Nat. Commun.* **14**, 4334 (2023).
792 36. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
793 37. Cingolani, P. *et al.* A program for annotating and predicting the effects of single
794 nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster
795 strain w1118; iso-2; iso-3: SNPs in the genome of Drosophila melanogaster strain w1118;
796 iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
797 38. Korani, W. *et al.* De novo QTL-seq Identifies Loci Linked to Blanchability in Peanut
798 (Arachis hypogaea) and Refines Previously Identified QTL with Low Coverage Sequence.
799 *Agronomy (Basel)* **11**, (2021).
800 39. Davies, R. W., Flint, J., Myers, S. & Mott, R. Rapid genotype imputation from sequence
801 without reference panels. *Nat. Genet.* **48**, 965–969 (2016).
802 40. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic
803 features. *Bioinformatics* **26**, 841–842 (2010).
804 41. Johnson, M. G. *et al.* HybPiper: Extracting coding sequence and introns for phylogenetics
805 from high-throughput sequencing reads using target enrichment. *Appl. Plant Sci.* **4**,
806 1600016 (2016).
807 42. Katoh, K., Misawa, K., Kuma, K.-I. & Miyata, T. MAFFT: a novel method for rapid
808 multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**,
809 3059–3066 (2002).
810 43. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated
811 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973
812 (2009).
813 44. Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic
814 inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
815 45. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S.
816 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**,
817 587–589 (2017).
818 46. Junier, T. & Zdobnov, E. M. The Newick utilities: high-throughput phylogenetic tree

processing in the UNIX shell. *Bioinformatics* **26**, 1669–1670 (2010).
47. Mirarab, S. *et al.* ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541-8 (2014).
48. Rambaut, A. *FigTree. A Graphical Viewer of Phylogenetic Trees*. (2018).
49. Bradbury, P. J. *et al.* TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
50. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer, New York, NY, 2009).
51. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. Ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
52. Caye, K., Deist, T. M., Martins, H., Michel, O. & François, O. TESS3: fast inference of spatial population structure and genome scans for selection. *Mol. Ecol. Resour.* **16**, 540–548 (2016).
53. Hijmans, R. J., Barbosa, M., Ghosh, A. & Mandel, A. *Geodata*. (2024).
54. Hijmans, R. J. *et al. Terra*. (2025).
55. Hijmans, R. J., van Etten, J., Sumner, M., Cheng, J. & Baston, D. *Raster*. (2025).
56. Thornthwaite, C. W. An Approach toward a Rational Classification of Climate. *Geogr. Rev.* **38**, 55–94 (1948).
57. López-Hernández, F. & Cortés, A. J. Last-generation genome-environment associations reveal the genetic basis of heat tolerance in common bean (Phaseolus vulgaris L.). *Front. Genet.* **10**, 954 (2019).
58. Zomer, R. J., Xu, J. & Trabucco, A. Version 3 of the Global Aridity Index and Potential Evapotranspiration Database. *Sci. Data* **9**, 409 (2022).
59. Lipka, A. E. *et al.* GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28**, 2397–2399 (2012).
60. Wang, J. & Zhang, Z. GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. *Genomics Proteomics Bioinformatics* **19**, 629–640 (2021).
61. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
62. Nyquist, W. E. & Baker, R. J. Estimation of heritability and prediction of selection response in plant populations. *CRC Crit. Rev. Plant Sci.* **10**, 235–322 (1991).
63. R Core Team. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. (Vienna, Austria, 2021).

64. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
65. Fisher, T. R. & Cruden, R. Chromosome Number and Observations in the Genus Silphium. *Ohio J. Sci.* **62**, 258–259 (1962).
66. Settle, W. J. The Chromosome Morphology in the Genus Silphium. *Ohio J. Sci.* **67**, 10–19 (1967).
67. Zeng, P. *et al.* Two Melanthiaceae genomes with dramatic size difference provide insights into giant genome evolution and maintenance. *Nat. Plants* **11**, 1500–1513 (2025).
68. Settle, R. & Fisher, W. J. The varieties of Silphium integrifolium. *Rhodora* **72**, 536–543 (1970).
69. Raduski, A. R. *et al.* Patterns of genetic variation in a prairie wildflower, Silphium integrifolium, suggest a non-prairie origin and locally adaptive variation. *Am. J. Bot.* **108**, 145–158 (2021).
70. Norris, E. T. *et al.* Admixture-enabled selection for rapid adaptive evolution in the Americas. *Genome Biol.* **21**, 29 (2020).
71. Benevenuto, J., Ferrão, L. F. V., Amadeu, R. R. & Munoz, P. How can a high-quality genome assembly help plant breeders? *Gigascience* **8**, (2019).
72. Vilela, A. *et al.* Progress and Bottlenecks in the Early Domestication of the Perennial Oilseed Silphium integrifolium, a Sunflower Substitute. *Sustain. Sci. Pract. Policy* **10**, 638 (2018).
73. Prasifka, J. R., Peterson, K., Mallinger, R. E. & Van Tassel, D. L. Changes to architecture of *Silphium integrifolium* Michx. during domestication reveal new trade-offs for yield. *Crop Sci.* **62**, 1060–1068 (2022).
74. Xiang, Y.-H. *et al.* An α/β hydrolase family member negatively regulates salt tolerance but promotes flowering through three distinct functions in rice. *Mol. Plant* **15**, 1908–1930 (2022).
75. Zhang, H. *et al.* A DTX/MATE-type transporter facilitates abscisic acid efflux and modulates ABA sensitivity and drought tolerance in Arabidopsis. *Mol. Plant* **7**, 1522–1532 (2014).
76. Hsieh, M.-H. & Goodman, H. M. Molecular characterization of a novel gene family encoding ACT domain repeat proteins in Arabidopsis. *Plant Physiol.* **130**, 1797–1806 (2002).
77. Hyten, D. L. *et al.* Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 16666–16671 (2006).

78. Crain, J., Haghighattalab, A., DeHaan, L. & Poland, J. Development of whole-genome prediction models to increase the rate of genetic gain in intermediate wheatgrass (Thinopyrum intermedium) breeding. *Plant Genome* **14**, e20089 (2021).
79. Brazier, T. & Glémin, S. Diversity and determinants of recombination landscapes in flowering plants. *PLoS Genet.* **18**, e1010141 (2022).
80. Mascher, M. *et al.* A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**, 427–433 (2017).
81. Rodgers-Melnick, E. *et al.* Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 3823–3828 (2015).
82. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).

6 Figures

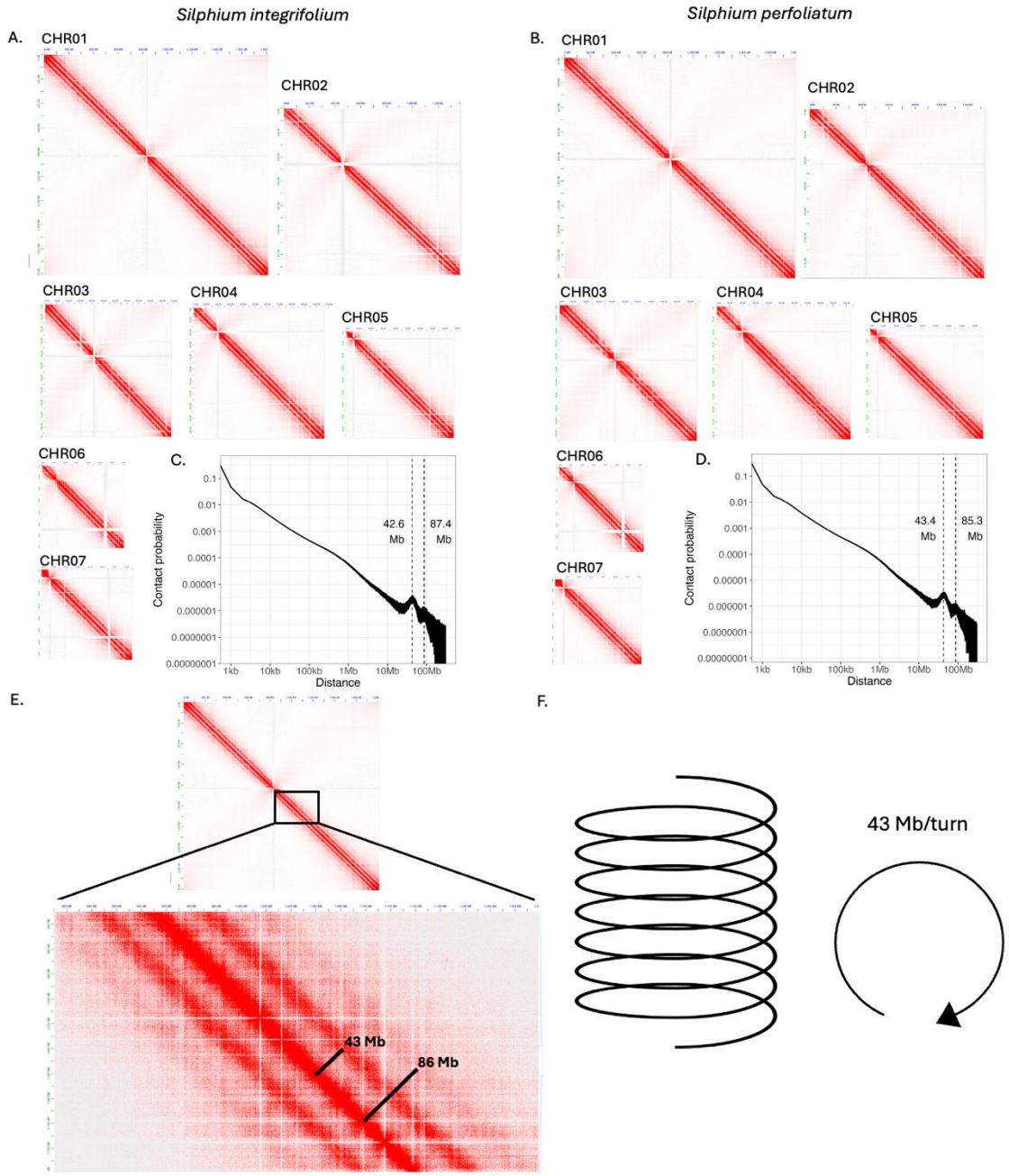

Figure 1. Genome assembly for *Silphium integrifolium* and *S perfoliatum*. A. Contact maps for the 7 chromosomes at scale in *Silphium integrifolium*. B. Contact maps for the 7 chromosomes at scale in *Silphium perfoliatum*. Probability of contact based on the fraction of paired end reads that are found at different distances from one another in 1kb windows in *S. integrifolium* (C.) and *S. perfoliatum* (D.). Information of paired-end reads that are found at different distances from one another were obtained from a mapped Omni-C file using the tool HOMER[82]. Axis units are presented in a log10 scale. The dashed lines show the peaks for interaction at 42.65 Mb and 43.39 Mb for *S. integrifolium* and *S. perfoliatum*, respectively. E. Silphium integrifolium chromosome 1 (850Mb to 1240Mb) in detail. The distance between the parallel diagonals is the average between the values for the peaks for *S. integrifolium* and *S. perfoliatum*. F. Proposed coil structure of chromatin during interphase.

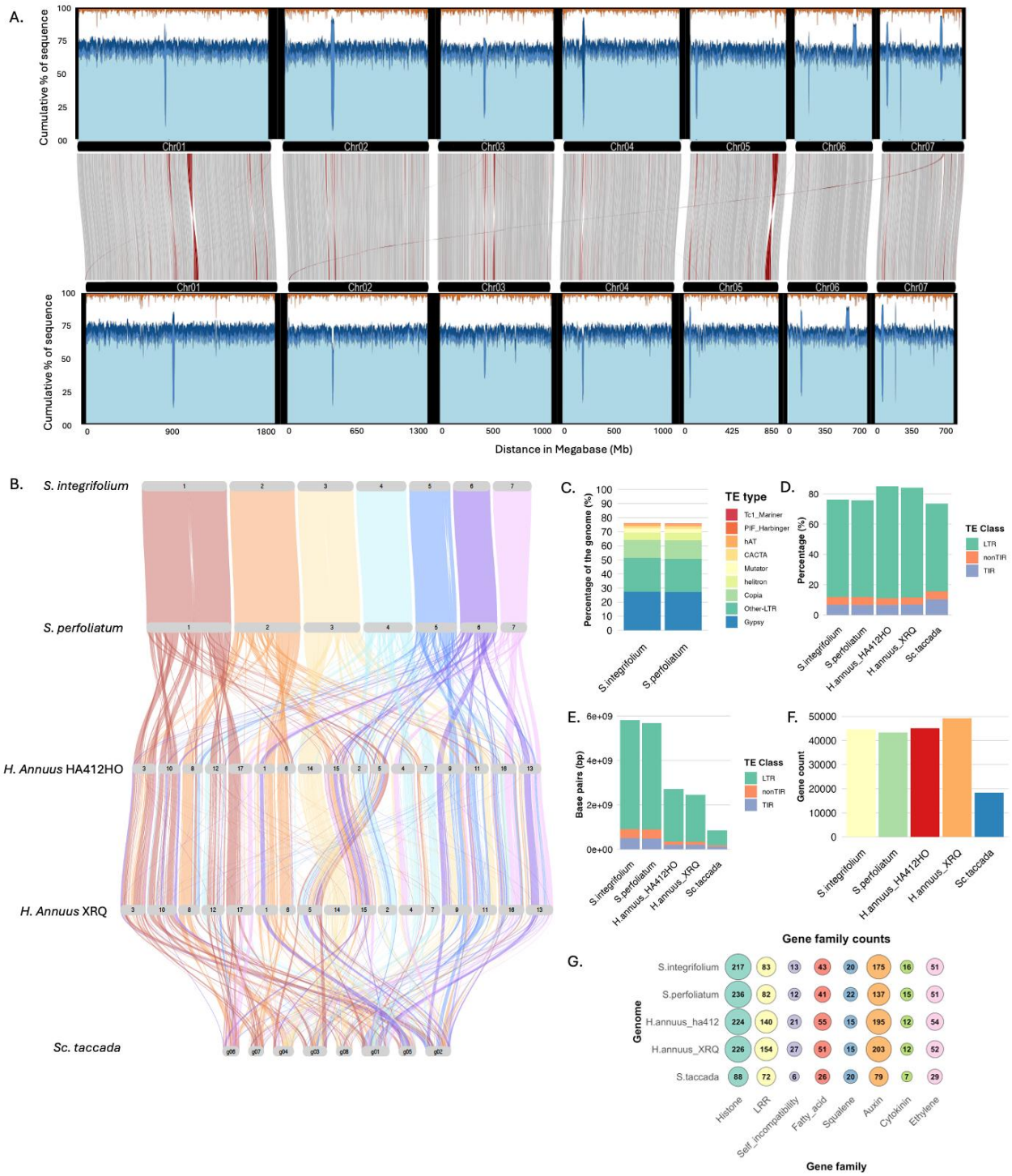

Figure 2. Comparative genomics of *Silphium integrifolium* and *S. perfoliatum* in the context of the Asteraceae. A. Annotation and synteny between *S. integrifolium* (top) and *S. perfoliatum* (bottom) (LTR - light blue, TIR - blue, Helitron - dark blue), genes (orange), unknown sequences (white). Cumulative proportions of annotation calculated in 5Mb windows with 1Mb step. B. Synteny among the two *Silphium* genomes, two Sunflower genomes and *Scaevola taccada*. Synteny plots generated with genespace [27]. C. Percentage of each type of transposable element in the Silphium genomes. D. Percentage of different classes of transposable elements in the five genomes analyzed. E. Amount in base pairs of transposable elements in among the five genomes analyzed. F. Total gene number across the genomes and G. Number of genes in different gene families.

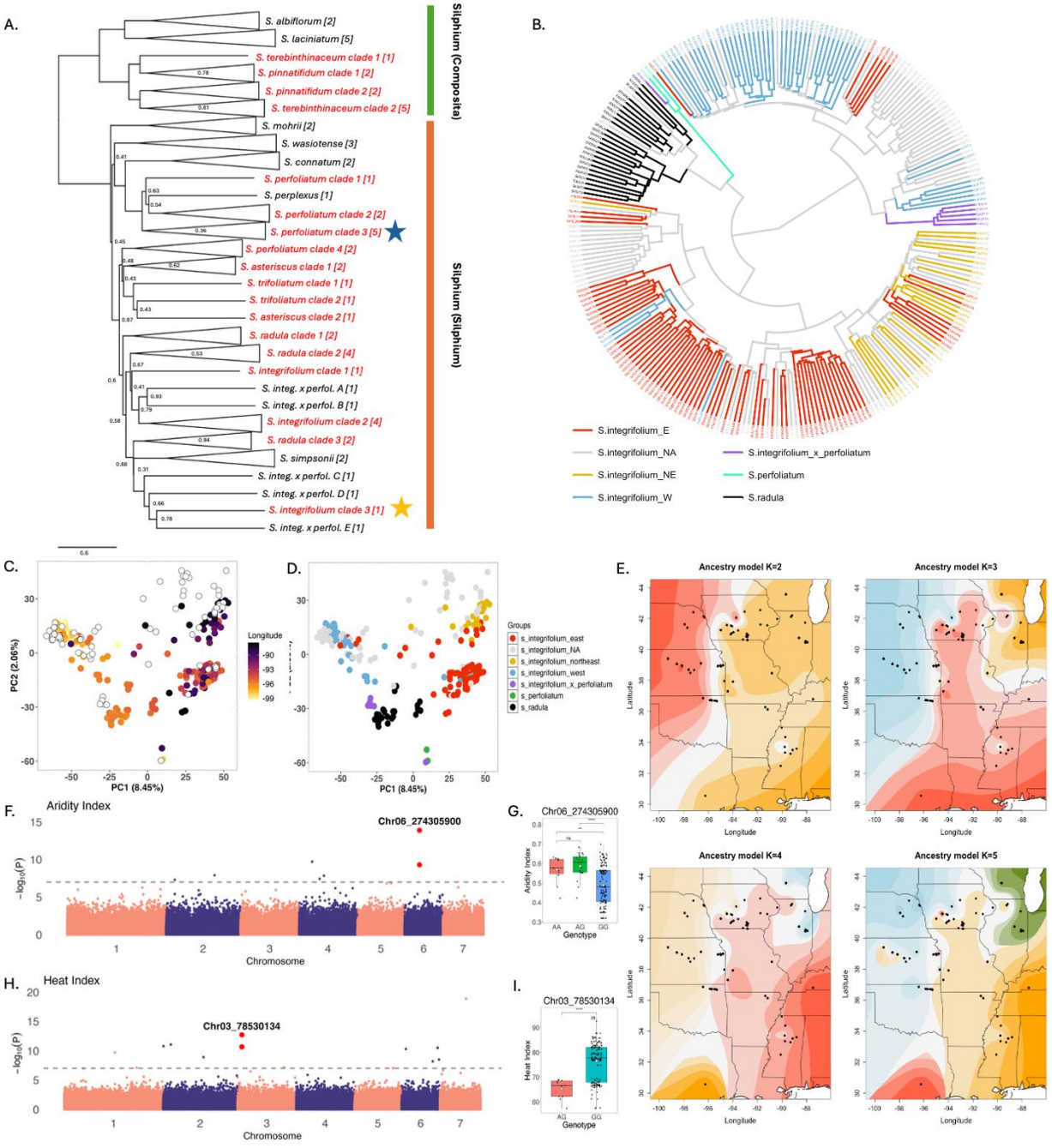

Figure 3. Phylogeny, population structure, diversity and environmental adaptation in *Silphium*. A. Phylogeny of *Silphium* including 14 species; polyphyletic species shown in red text. Blue and yellow stars indicate the position of the "*1JHW*" and "*Bad Astra*" in the phylogeny. See complete tree, including outgroups, in Supplementary Figure S8. B. Hierarchical clustering of the *Silphium* diversity panel (N=258) based on the kinship matrix calculated with 449,554 SNPs. Individuals with known GPS coordinates were colored according to the group from the spatial ancestry (East - Red, West - Blue, Northeast - Orange, *S. radula* - Black). C-D: Population genomic principal component analysis PC1 and PC2 plots of the Silphium association panel. C. color layer based on longitude (N=183 georeferenced individuals). D. color layer based on the groups from the hierarchical cluster. E. Spatial ancestry of 156 georeferenced *S. integrifolium* individuals with clustering set from 2 to 5. F. Genome environment analysis for aridity index. G. Effect of SNP Chr06_27430590 on the aridity index. H. Genome environment analysis for heat index. I. Effect of the SNP Chr03_78530134 on the heat index.

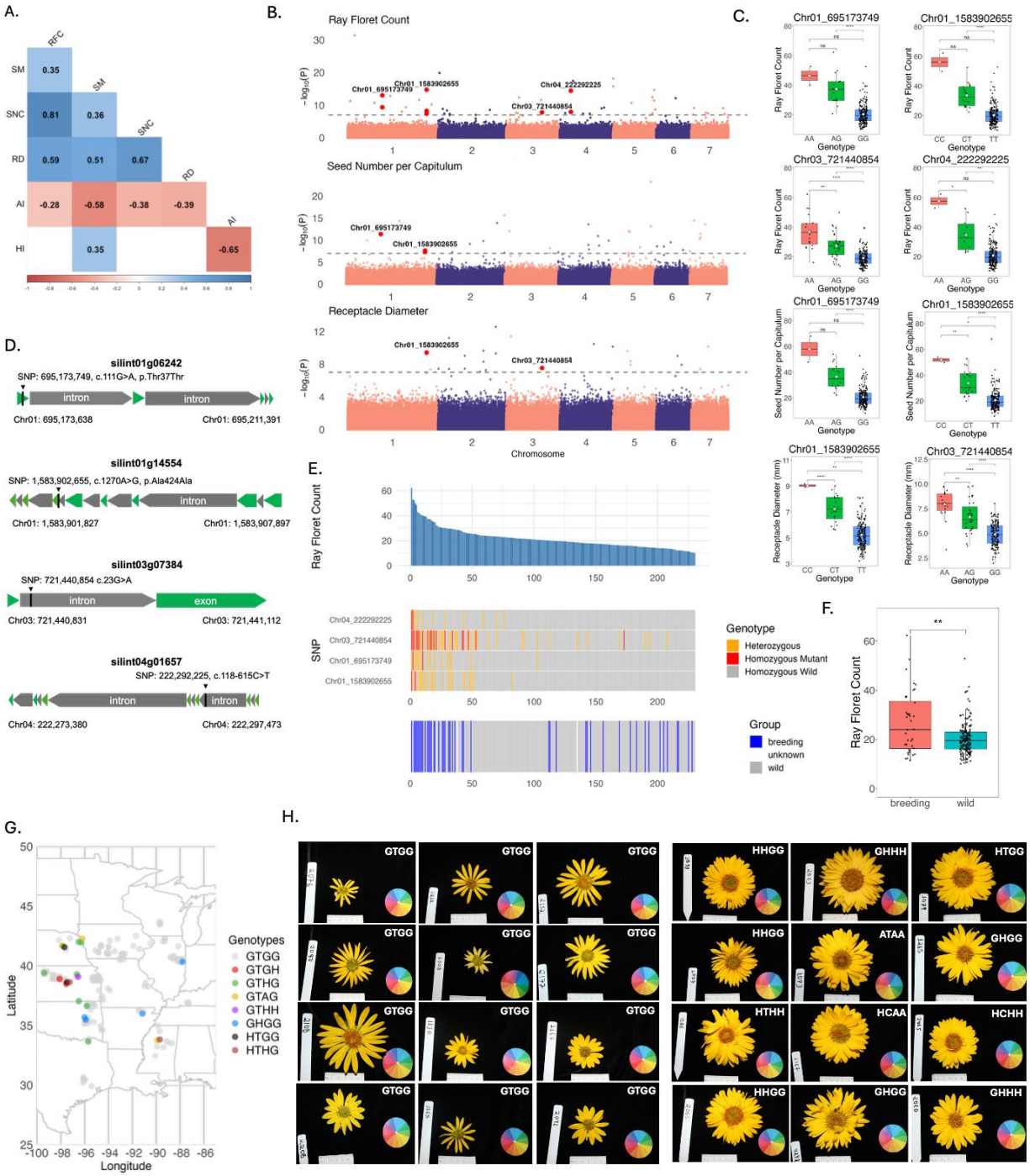

Figure 4. GWAS for domestication traits in *Silphium*. A. Pearson correlation based on BLUE values of the combined analysis in 2025 for the traits studied. Only significant correlations p<0.05 are shown. B. Manhattan plots with the combined results from two models (BLINK and FarmCPU) for ray floret count (RFC), seed number per capitulum (SNC) and receptacle diameter (RD) with loci explaining > 20 % of variance highlighted. Dotted horizontal lines represent the Bonferroni threshold -log(p)=7. C. Effect of the different allele combinations on the phenotypes (BLUEs). D. gene models where significant SNPs were detected. The models are not represented at scale. E. Ray floret count ordered from highest to lowest, genotypes at the four loci corresponding to the individuals ordered and groups (breeding vs wild) with the individuals. F. Ray floret count in breeding vs wild genets. G. Origin of the mutant genotypes. H. Effect of the variants on the ray floret count.

7 Tables

Table 1. Genome assembly statistics.

	Bad Astra	1JHW
Total assembly size (bp)	7641408655	7507277660
Total number of genes	44653	43390
Number of contigs	332	241
N50 contig length (Mb)	41.3	53.9
Longest contig (Mb)	115	183
Number of scaffolds	7	7
Complete BUSCOs (%)	97.5	98
LTR Assembly Index	14.46	14.86

Table 2. Transposable element content in *S. integrifolium* (Bad Astra) and *S. perfoliatum* (1JHW).

	Bad Astra		**1JHW**	
TE Class	Count	Proportion (%)	Count	Proportion (%)
Long Terminal Repeats (LTRs)				
Copia	787126	12.83	818048	13.18
Gypsy	2253310	27.40	2136211	27.20
unknown	2744810	23.92	2621003	23.48
Terminal Inverted Repeats (TIRs)				
CACTA	322868	1.67	310105	1.64
Mutator	545382	2.90	543906	2.77
PIF_Harbinger	82054	0.34	85614	0.37
Tc1_Mariner	16587	0.06	22067	0.09
hAT	382142	1.61	367325	1.57
non-TIR				
helitron	1258302	5.33	1270613	5.43
Total		**76.06**		**75.73**

1003 Table 3. GWAS loci detected in both models and multiple environments explaining more than 10% of the phenotypic variance.

SNP	Chr	Position	P.value	MAF	Model	Trait*	Environment**	N	Alleles	PVE	Gene	Annotation
Chr01_1583902655	1	1583902655	5.53862E-09	0.05	BLINK	RFC	Combined	230	T/C	29.39	silint01g14554	MATE family efflux transporter
			3.98176E-08	0.05	FarmCPU	RFC	Combined	230	T/C	29.39		
			1.8178E-15	0.05	BLINK	RFC	KS24	190	T/C	29.39		
			3.84037E-10	0.06	FarmCPU	RD	AL25	194	T/C	24.30		
			2.28687E-08	0.05	BLINK	SNC	KS25	198	T/C	25.01		
			4.86223E-08	0.05	FarmCPU	SNC	KS25	198	T/C	25.01		
Chr01_695173749	1	695173749	4.22848E-10	0.04	BLINK	RFC	Combined	230	G/A	22.82	silint01g06242	Alpha/beta hydrolase
			9.81729E-14	0.04	FarmCPU	RFC	Combined	230	G/A	22.82		
			4.17964E-12	0.05	FarmCPU	SNC	AL25	194	G/A	23.37		
Chr02_1160355565	2	1160355565	4.93281E-10	0.19	BLINK	RD	Combined	225	A/T	11.22	silint02g11897	Putative transcription factor FAR
			2.61733E-13	0.19	FarmCPU	RD	Combined	225	A/T	11.22		
Chr03_721440854	3	721440854	1.60791E-08	0.17	BLINK	RFC	AL25	196	G/A	31.20	silint03g07384	Unknown protein
			3.00152E-08	0.16	FarmCPU	RD	KS25	199	G/A	28.48		
Chr03_78530134	3	78530134	2.14615E-11	0.03	BLINK	HI	Index	183	G/A	12.45	intergenic window	-
			1.93344E-13	0.03	FarmCPU	HI	Index	183	G/A	12.45		
Chr03_910576785	3	910576785	9.04611E-08	0.06	BLINK	SNC	Combined	225	A/G	18.28	silint03g09429	DDB1 and CUL4 associated factor 4
			7.13689E-09	0.06	FarmCPU	SNC	Combined	225	A/G	18.28		
Chr04_222292225	4	222292225	3.72995E-15	0.03	BLINK	RFC	KS25	206	G/A	23.03	silint04g01657	ACT domain containing protein ACR4
			1.15077E-08	0.03	FarmCPU	RFC	KS25	206	G/A	23.03		
Chr04_603756119	4	603756119	1.24541E-09	0.03	BLINK	SNC	KS25	198	T/A	18.31	intergenic window	-
			2.75326E-14	0.03	FarmCPU	SNC	KS25	198	T/A	18.31		
Chr05_11691107	5	11691107	7.51674E-19	0.02	BLINK	RFC	Combined	230	G/A	14.38	intergenic window	-
			1.71952E-16	0.02	FarmCPU	RFC	Combined	230	G/A	14.38		
Chr05_794020568	5	794020568	2.77243E-08	0.03	BLINK	RFC	AL25	196	A/C	12.80	silint05g07490 (10Kb downstream)	Thioredoxin-like protein
			3.73566E-13	0.03	FarmCPU	RFC	AL25	196	A/C	12.80		
			5.185E-24	0.02	FarmCPU	SNC	AL25	194	A/C	13.89		
Chr05_831132026	5	831132026	2.25069E-15	0.02	BLINK	RFC	KS25	206	G/A	18.35	silint05g07840 (50Kb upstream)	Unknown protein
			3.79352E-12	0.02	FarmCPU	RFC	KS25	206	G/A	18.35		
Chr06_274305900	6	274305900	5.18447E-10	0.12	BLINK	AI	Index	183	G/A	12.75	silint06g02768 (50Kb upstream)	ALP1-like protein
			1.2544E-14	0.12	FarmCPU	AI	Index	183	G/A	12.75		

* Traits extracted from geospatial databases, AI: Aridity Index; HI: Heat Index. Traits measured in field-grown plants, RD: Receptacle Diameter; RFC: Ray Floret Count, SNC: Seed Number per Capitulum.

**Environments, AL24: Alabama 2024; AL25: Alabama 2025, KS24: Kansas 2024; KS25: Kansas 2025; Combined: all 4 environments; Index: environmental indices based on multiyear average environments from the location where genet originated.

Other Abbreviations: SNP: Single Nucleotide Polymorphism; Chr: Chromosome; MAF: Minor Allele Frequency; N: Number of samples; PVE: Phenotypic variance explained

1009

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- silphiumsupplementaryfilefinal.xlsx
- silphiumsupplementaryfiguresandtablesfinal.docx
- sintegrifolium.ann.5.7617700679632.zip