nature communications

Article

https://doi.org/10.1038/s41467-026-75205-3

Assembly of *Silphium* interspecific hybrid genomes opens the genus to phylogenomics, ecogenomics, and molecular breeding

Received: 31 January 2026

Accepted: 25 June 2026

Published online: 09 July 2026

Check for updates

Renan S. Souza[1], Josh P. Clevenger[1], Jerry Jenkins[1], Walid Korani[2], Jeremy Schmutz[1], Jane Grimwood[1], Christopher Plott[1], Yaniv Brandvain[3], Kathryn Turner[4], Allison J. Miller[5,6], Brent S. Hulke[7], Louis F. Nastasi[8], Ebony G. Murrell[9], Lori H. Handley[1], Jenell Webber[1], Lori Beth Boston[1], Stella Woeltjen[5], Matthew J. Rubin[6], Betsy Trana[4], Erin Howard[10], Ashlyn Bazzel[11], Alex Harkess[1] & David L. Van Tassel[4]

Wild perennial plants can be domesticated to make agriculture more diverse and resilient, but many have large genomes that have been recalcitrant to analysis. Here, we report phased genome assemblies for *Silphium integrifolium* Michx. and *S. perfoliatum* L., two species native to North America under domestication, and demonstrate the utility of trio-binning for genome assembly using an interspecific hybrid. These genomes have chromosomes reaching 1.8 Gb and a helical structure preserved during interphase with a loop circumference of 43 Mb. A genome-informed low coverage and target sequencing strategy enables the refinement of the genus phylogeny, reveals the spatial distribution and structure of natural populations, and identifies 81 loci associated with environmental and domestication traits. Variants in a MATE transporter, α/β hydrolase, and ortholog of Arabidopsis ACT Domain Repeat (ACR4) protein explain significant variance in floral architecture. These advances in genome assembly and genotyping could expand the range of candidates for de novo crop domestication.

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 plant species that can be used as food sources but are not actively cultivated[2]. Among these wild candidates, perennial species hold promise and value, as they may provide an avenue to marry high productivity, lower agricultural inputs, and ecosystem sustainability[3].

Traditional domestication paths took millennia and the intense selection for germination, increased yield and harvestability in specific environments, has led to the unintended loss of favorable alleles for other traits such as nutritional value and stress resistance[4]. Modern development of crops from wild plants has been rare, but domestication and breeding efforts should be greatly accelerated by recent advances in genomics and breeding methods[5]. These advances have made possible the rapid domestication of wild or semi-wild

[1]HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA. [2]Veil Genomics, Huntsville, AL, USA. [3]University of Minnesota, Minneapolis, MN, USA. [4]The Land Institute, Salina, KS, USA. [5]The Donald Danforth Plant Science Center, Saint Louis, MO, USA. [6]Saint Louis University, Saint Louis, MO, USA. [7]USDA-Agricultural Research Service, Edward T Schafer Agricultural Research Center, Fargo, ND, USA. [8]Department of Entomology, University of Nebraska-Lincoln, Lincoln, NE, USA. [9]Savanna Institute, Madison, WI, USA. [10]The University of Georgia, Athens, GA, USA. [11]Clemson University, Clemson, SC, USA.
e-mail: vantassel@landinstitute.org

plants, such as wild allotetraploid rice[6]. This process, also known as de novo domestication, is characterized by the assembly of reference genomes, trait mapping, marker assisted selection, gene editing and the use of predictive models to accelerate the improvement of candidate crops[7]. These genomic methods can 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 traditional 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[8], most of which reproduce by outcrossing, including interspecific hybridization[9]. *S. integrifolium* Michx. and *S. perfoliatum* L. are currently undergoing domestication, with the first showing promise as an oilseed crop with high tolerance to drought[10] and the second as a biomass and fiber crop[11]. *S. integrifolium* seeds have considerable concentration of oil (11.8 to 25.3 % w/w)[12] and high concentrations of squalene (4.9% of the oil), a triterpene used as an emollient in cosmetics and as an adjuvant in vaccines[13]. Squalene has traditionally been sourced from shark livers, but in light of conservation efforts, producing this compound from plants is of great interest[14].

In addition to their potential as oilseed and biomass crops, *Silphium* species play key ecological roles in the prairie ecosystem[15]. 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 periods of drought. 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[16,17].

The lack of genetic studies and genomic resources for wild species with conservation and agricultural potential like *Silphium* underscores an urgent need to develop and deploy approaches to advance de novo domestication. One of the approaches to advance this effort is the survey of the genetic variation in natural populations and wild germplasm collections via population structure analysis, genetic relatedness and spatial ancestry. These analyses are common in model organisms and major crops, but they are still not used as a tool to survey candidates for domestication. This is largely because our ability to perform these studies depends on cost-effective genotyping methods, which are still not available for most plant species.

Developing resources such as a reference genome and genotyping methods are one of the initial steps to expand the characterization and utilization of genetic diversity in both basic and applied research[18]. Not many years ago, those of us working on minor crops or wild species believed that the cost and time required for these species to enter the genomics space would be prohibitive. Here, we show that integrating genomic technologies allows the high-quality assembly of large genomes and enables a range of genomic investigations. Genome assemblies and annotations provide the foundation for the development of low coverage short-read and targeted sequencing of wild relatives, wild germplasm accessions, and breeding materials. The genotyping of these populations increases the resolution in the phylogeny and elucidates the population structure and the diversity in the natural range of *Silphium* species. The integration of these data with phenotypes and environmental indexes reveals the genomic architecture of domestication traits and adaptation and offers the most comprehensive genomic analysis for the genus to date, providing critical genomic resources to support genomics-enabled domestication and restoration efforts.

Table 1 | Genome assembly statistics

	Bad Astra	1JHW
Total assembly size (bp)	7,641,408,655	7,507,277,660
Total number of genes	44,653	43,390
Number of contigs	332	241
N50 contig length (Mb)	41.30	53.90
Longest contig (Mb)	115	183
Number of scaffolds	7	7
Complete BUSCOs (%)	97.50	98
LTR Assembly Index	14.46	14.86

Results

Silphium genomes exhibit a putative coil structure and are mainly composed of LTRs

The genome assembly resulted in 332 contigs for *S. integrifolium* 'Bad Astra' and 241 contigs for *S. perfoliatum* '1JHW' and the scaffolding resulted in 7 chromosomes per haplotype. Contig N50 and final assembly size were 41.3 Mb and 7.6 Gb for 'Bad Astra' and 54.7 Mb 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 'Bad Astra' and 14.86 for '1JHW', indicating high-quality assemblies. Both *Silphium* genomes are composed of seven pairs of long chromosomes, with the length of chromosome 1 reaching 1.81 Gb in *S. integrifolium* and 1.79 Gb in *S. perfoliatum* (Supplementary Data 1). The genomes of both species are highly heterozygous, with *S. integrifolium* and *S. perfoliatum* having 2.24% and 1.44% heterozygous loci, respectively (Supplementary Fig 1). The chloroplast genome assembly resulted in a sequence of 152,169 bp with 113 genes annotated for both genomes (Supplementary Fig 2).

When analyzing the Omni-C contact map, we observed the usual main diagonal indicating that each point along the DNA sequence makes the most contacts with the points flanking it (Fig. 1a, b). Unexpectedly, we observed a weaker pair of diagonals in parallel with the main diagonal (Fig. 1e) and an increase in chromatin contact at approximately 43 Mb (Fig. 1c, d). Another pair of even weaker diagonals was discernible, each tertiary diagonal approximately 43 Mb outside of the secondary diagonal (Fig. 1c, d). 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 chromatin in *Silphium* has a tertiary structure: a helix with 43 Mb per turn (Fig. 1f) that is maintained in both *S. integrifolium* and *S. perfoliatum* genomes during interphase.

To contextualize the *Silphium* genomes within the Asteraceae family, we analyzed the patterns of synteny among the two *Silphium* genomes, two *Helianthus annuus* genomes and the outgroup species *Scaevola taccada* (Supplementary Note 1). Despite strong collinearity between *S. integrifolium* and *S. perfoliatum*, there were two major inversions, one on chromosome 1 spanning 44.64 Mb (1057.31 Mb to 1101.95 Mb) and another on chromosome 5 spanning 51.02 Mb (745.83 Mb to 796.85 Mb) (Fig. 2a, Supplementary Fig 3). 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* (Fig. 2a). LTR elements predominated in all genomes compared, and in *S. integrifolium*, LTRs corresponded to 64.1% of the genome and the transposons in the Ty3/Gypsy family were the most common (27.4%) (Fig. 2b–d). In *S. perfoliatum*, LTR accounted for 63.9 % of the genome

https://doi.org/10.1038/s41467-026-75205-3

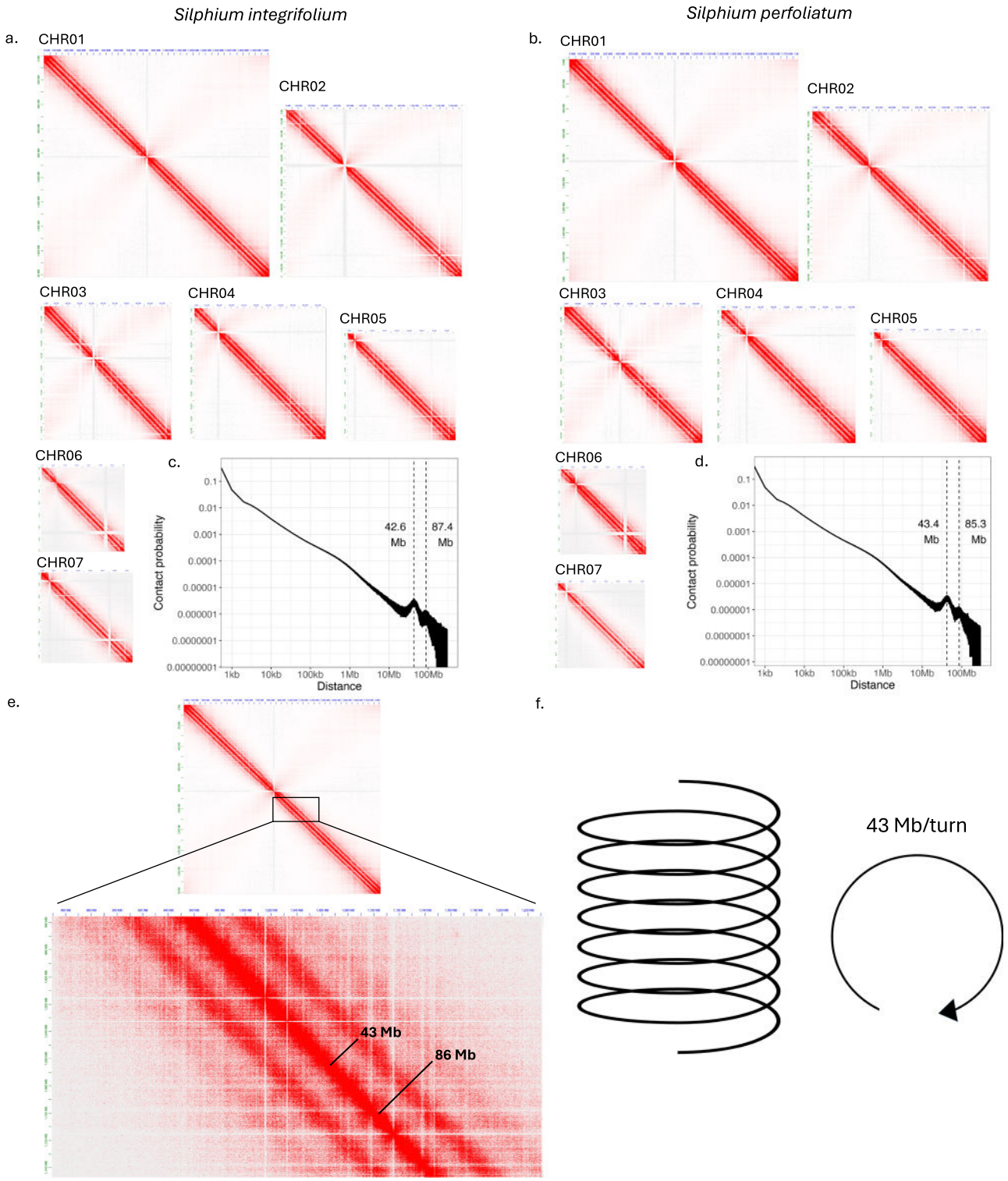

Fig. 1 | Genome assembly for *S. integrifolium* and *S. perfoliatum*. **a** Contact maps for the 7 chromosomes at scale in *S. integrifolium*. **b** Contact maps for the 7 chromosomes at scale in *S. perfoliatum*. Probability of contact based on the fraction of paired end reads that are found at different distances from one another in 1 kb 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[49]. 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** *S. integrifolium* chromosome 1 (850–1240 Mb) 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. Source data is provided as a Source Data file.

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 (94%) of this difference is due to repetitive elements (Supplementary Table 1).

Targeted sequencing recovers *Silphium* phylogenetic backbone and shows discrepancies

We reconstructed the phylogeny of *Silphium* aiming to resolve the relationship between 14 species using a set of 789 loci (Fig. 3a).

https://doi.org/10.1038/s41467-026-75205-3

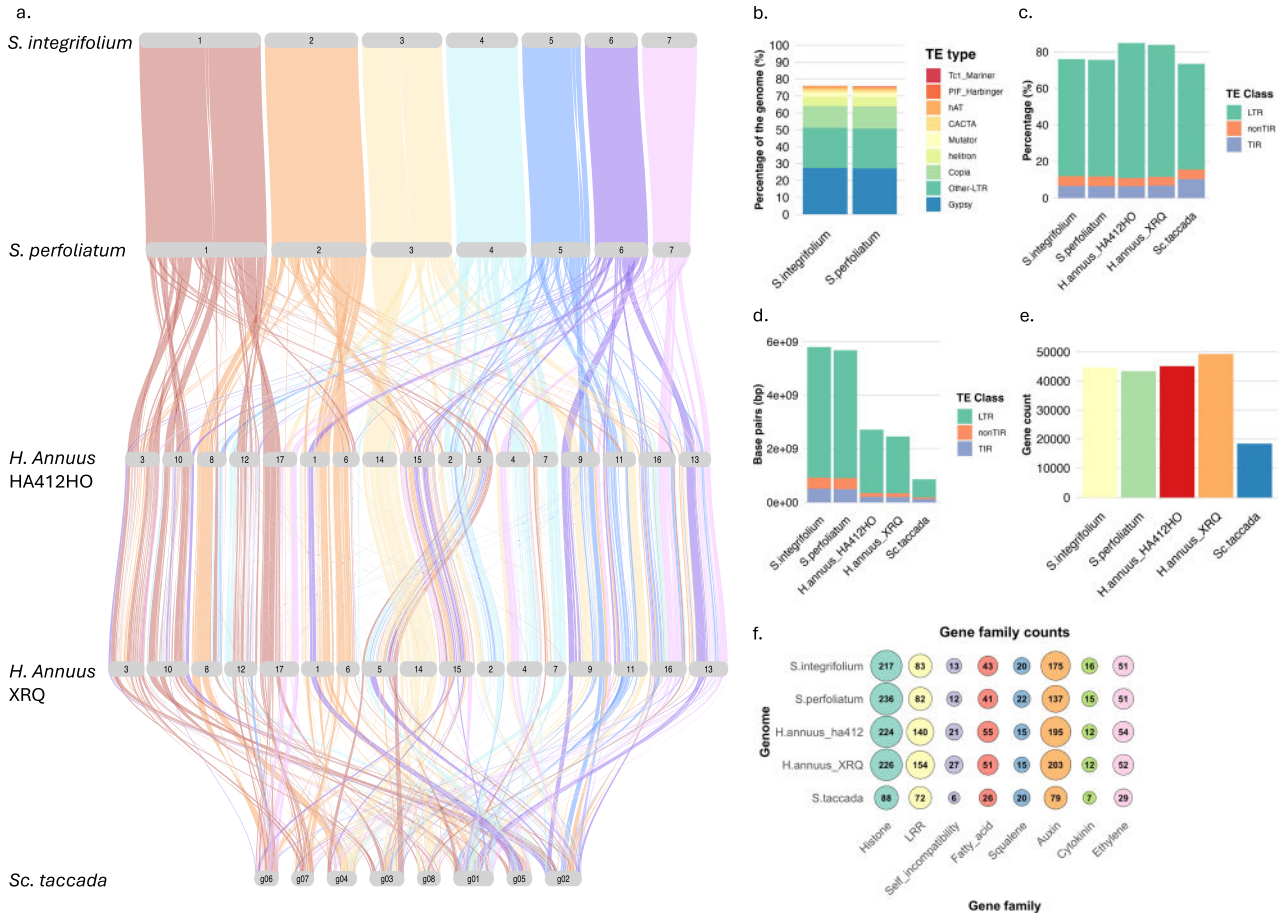

Fig. 2 | Comparative genomics of *S. integrifolium* and *S. perfoliatum* in the context of the Asteraceae. a Synteny among the two *Silphium* genomes, two sunflower genomes and *Scaevola taccada*. Synteny plots generated with genespace[53]. **b** Percentage of each type of transposable element in the *Silphium* genomes. **c** Percentage of different classes of transposable elements in the five genomes analyzed. **d** Transposable elements amount (base pairs) among the five genomes analyzed. **e** Total gene number across the genomes and **f** Number of genes in different gene families. Source data is provided as a Source Data file.

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 (Fig. 3a). *S. 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 Fig 4).

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 Fig 4). 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*. *S. asteriscus* was also polyphyletic with regard to *S. trifoliatum*, although *S. trifoliatum* is often considered a variety or subspecies of *S. asteriscus* because of ongoing taxonomic debate[8].

Silphium populations are divided into two major groups in its native range

The hierarchical clustering based on the kinship of all 258 genets in the *Silphium* association panel (SAP) indicated a top-level division of the population in two major groups (Fig. 3b). The PCA plot further corroborates the hypothesis of two populations when the points were color coded based on longitude and it became evident that genetic relatedness is strongly correlated with population location in the east-west gradient (Fig. 3c). The spatial ancestry analysis, using the subset of 156 *S. integrifolium* individuals from wild populations with known GPS coordinates, considered scenarios with 2 to 5 hypothesized ancestral populations (K) (Fig. 3e). With $K = 2$, there is a clear east-west division that persists even when more ancestral populations are considered. 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 Fig. 5). In fact, the SAP exhibited a high level of heterozygosity, with 14.9% of the individuals showing a heterozygous call per locus on average across the 449,554 SNPs (Supplementary Note 2). 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 could be further divided into

https://doi.org/10.1038/s41467-026-75205-3

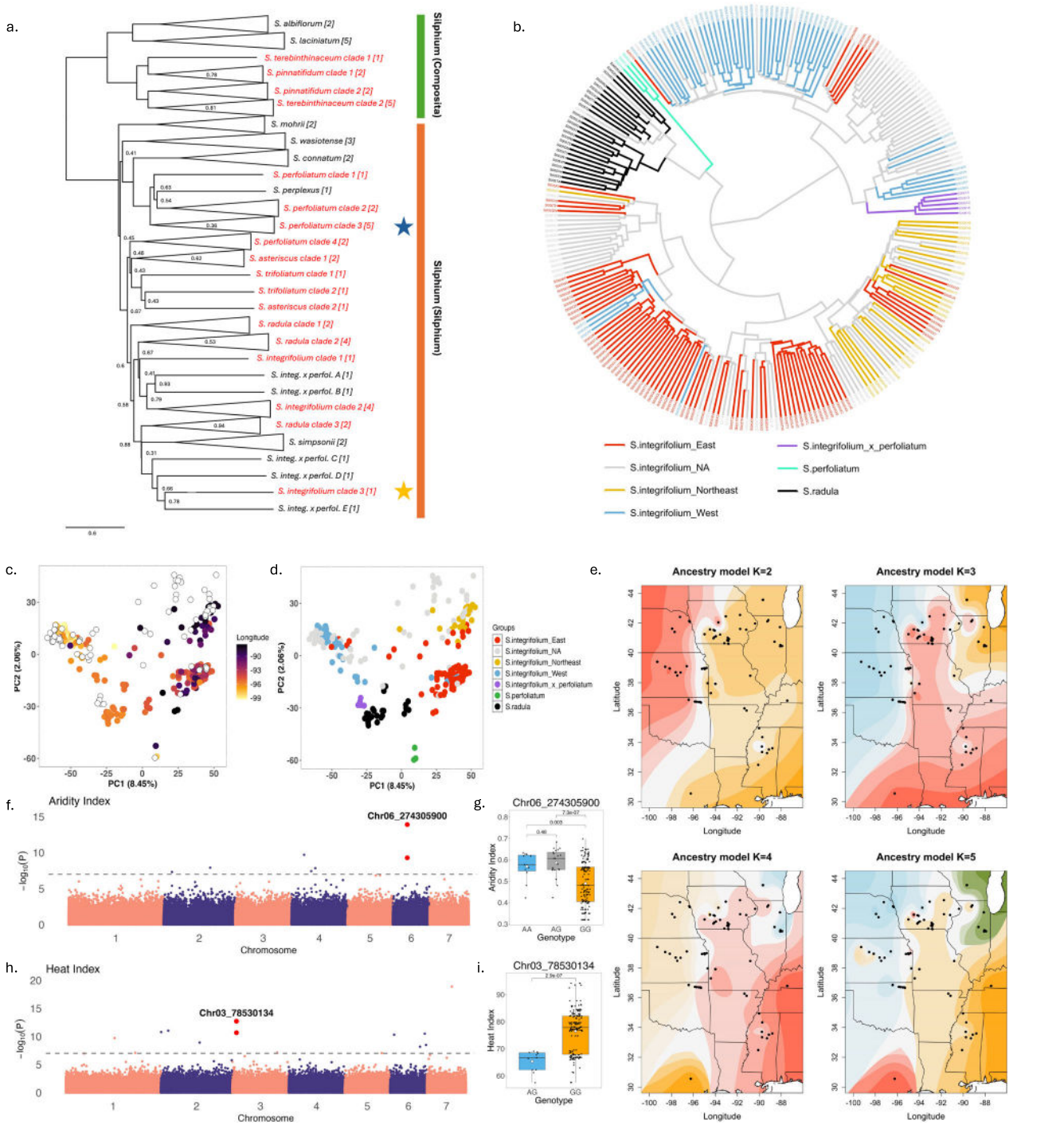

Fig. 3 | Phylogeny, population structure, diversity and environmental adaptation in *Silphium*. **a** Phylogeny of *Silphium* including 14 species; polyphyletic species shown in red text. Blue and yellow stars indicate the position of the '1JHW' and 'Bad Astra' in the phylogeny. See complete tree, including outgroups, in Supplementary Fig 4. **b** Hierarchical clustering of the *Silphium* association 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 model with K = 3 (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. Dotted horizontal lines represent the Bonferroni threshold -log(p)=7. **i** Effect of the SNP Chr03_78530134 on the heat index. In **f**, **h** significant markers are selected by combining iterative steps of Fixed Effect and Random Effect Model for FarmCPU and Bayesian Information Criterion for BLINK. Final *p*-values are calculated using single-marker least-squares regression. Comparisons presented in **g**, **i** are based on a two-tailed Student's *t* test with a total of 183 samples. In the boxplot, the center line represents the median, the lower and upper bounds of the box spans from the first to the third quartiles (the 25th and 75th percentiles), whiskers extend to 1.5 times the interquartile range, and points indicate outliers. Source data is provided as a Source Data file.

https://doi.org/10.1038/s41467-026-75205-3

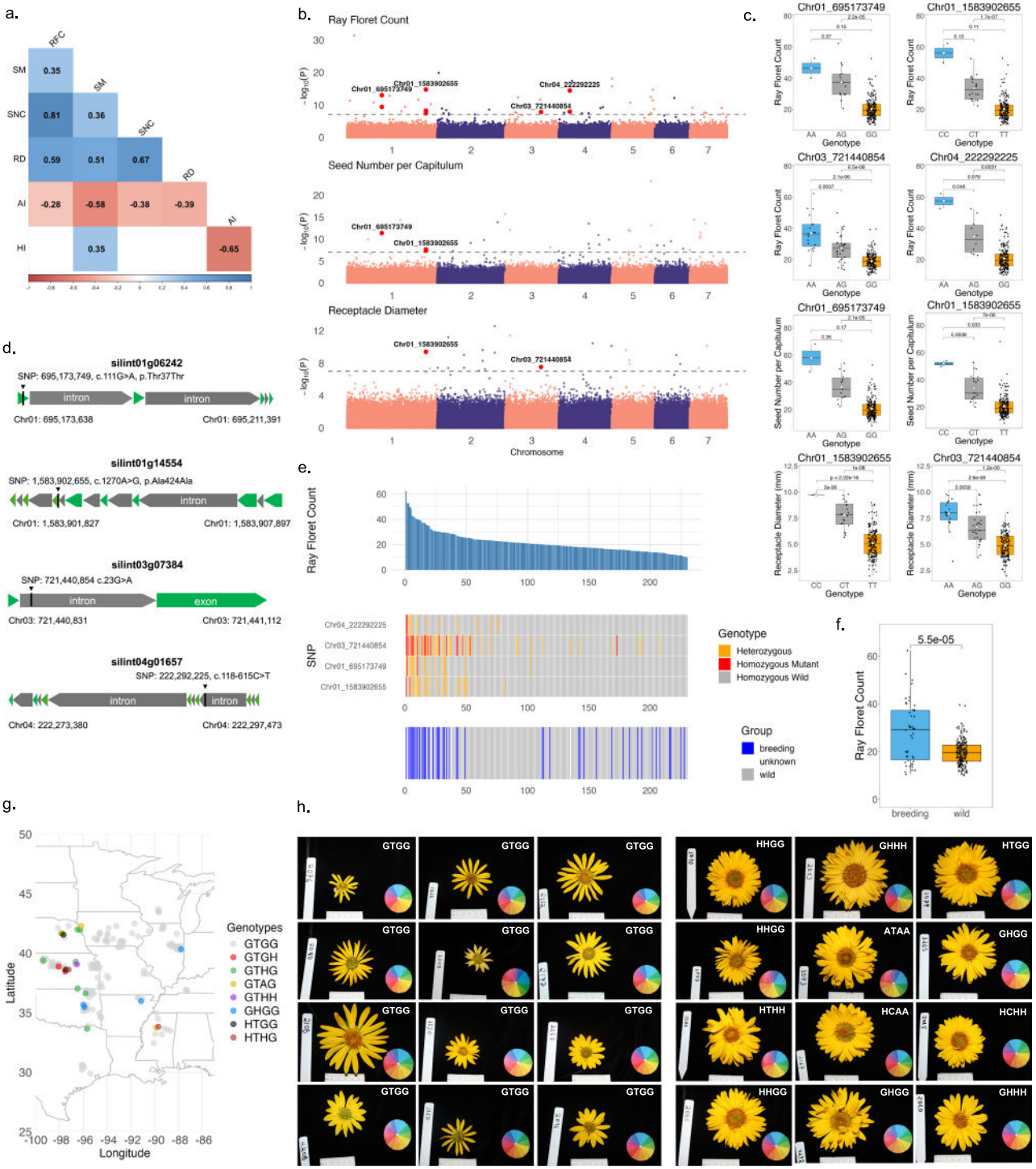

Fig. 4 | GWAS for domestication traits in *Silphium*. **a** Pearson correlation based on BLUE values of the combined analysis in 2025 for the traits studied. Only significant correlations $p < 0.05$ are shown. **b** Manhattan plots with the combined results from two models (BLINK and FarmCPU) for ray floret count (RFC), seed number per capitulum (SNC) and receptacle diameter (RD) with loci explaining >20% of variance highlighted in red. Dotted horizontal lines represent the Bonferroni threshold −log(p) = 7. Significant markers are selected by combining iterative steps of Fixed Effect and Random Effect Model for FarmCPU and Bayesian Information Criterion for BLINK. Final *p*-values are calculated using single-marker least-squares regression. **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** Genets ($N = 230$) ordered by ray floret count. Subplots show genet ray floret count ordered from high to low (multi-environment BLUEs), the genotype for the four domestication loci, and group identity (breeding vs. wild). **f** Ray floret count in breeding vs. wild genets. **g** Geographic distribution of the mutant genotypes. **h** Effect of the variants on the ray floret count. Comparisons presented on **c** ($N = 230$ for RFC and $N = 225$ for SNC and RD), and **f** ($N = 225$) are based on a two-tailed Student's *t* test. In the boxplot, the center line represents the median, the lower and upper bounds of the box spans from the first to the third quartiles (the 25th and 75th percentiles), whiskers extend to 1.5 times the interquartile range, and points indicate outliers. Source data is provided as a Source Data file.

four groups: *S. integrifolium* from the east, *S. integrifolium* from the west, *S. integrifolium* from northeast and *S. radula* (Fig. 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.

Genome-wide scans identify loci for adaptation and domestication traits

Four of the phenotypic traits, ray floret count (RFC), seed mass (SM), seed number per capitulum (SNC) and receptacle diameter (RD) that were evaluated in the SAP in four environments (AL24, AL25,KS24 and KS25) showed an overall high heritability, ranging from 0.74 to 0.92 (Supplementary Table 2). 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 with moisture availability in their source environment (the aridity index), with values ranging from −0.28 to −0.58 (Fig. 4a, Supplementary Fig. 6).

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%. From these 44, 12 loci were detected in both models and in multiple environments (Table 2, Supplementary Data 2, Supplementary Figs. 7–13). 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 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 (Fig. 3f–i).

For seed mass, 11 loci were identified, together explaining 56.04% of the variance. The SNP with the strongest association was Chr01_1030878591 with 9.03% of the variance for seed mass 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. 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 (Fig. 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% (Fig. 4b–d). For receptacle diameter, two SNPs described above 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 (Fig. 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.

When grouping the genets in the SAP into wild collections and breeding genets 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 the wild (AA genotype in SNP Chr03_721440854 in the accession S005T6) (Fig. 4e, Supplementary Data 3). Interestingly, there are three *S. radula* individuals, S005QM and 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 *S. integrifolium*. The four loci appear in the homozygous state for the favorable alleles more frequently in the breeding genets (Fig. 4e). This is also evident by comparing the floret count between wild and breeding genets (Fig. 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 (genotypes): the baseline wild type (GTGG) and seven mutant genotypes (GHGG, GTAG, GTGH, GTHG, GTHH, HTGG and HTHG). The mutant genotypes are present in several of the natural populations sampled, with the GTHG genotype 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 genotypes GTHH and GTGH occur only in genets collected in Kansas (Fig. 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 in the SAP (Fig. 4h).

Sensitivity analysis shows effect of population groups on the power to detect loci

To understand to what extent the different groups in the SAP and their domestication history impacted the power to detect the loci, we performed a sensitivity analysis by rerunning the genome scans with *S. integrifolium* individuals only ($N = 236$), wild individuals only ($N = 206$), and breeding individuals only ($N = 47$). We analyzed the effects of changes in the population size and allele frequencies on the *p*-value. In the comparison between full panel with the subsets, there was a significant change in statistical power for the 81 loci detected by genome scans. When considering only *S. integrifolium* genets, from the 81 loci detected in the entire population, only 12 remained significant, including three of the four capitulum loci, Chr01_1583902655 ($p = 3.61 \times 10^{-10}$), Chr01_695173749 ($p = 1.71 \times 10^{-33}$) and Chr03_721440854 ($p = 1.77 \times 10^{-08}$). In the wild only group, 13 loci remained significant, and all of them were related to environmental index, including the two major loci Chr03_78530134 ($p = 1.93 \times 10^{-13}$) and Chr06_274305900 (1.25×10^{-14}). In the breeding only group, none of the 81 loci were detected (Supplementary Fig 14). Overall, these results show a loss of power as populations become smaller.

Discussion

In this study, we report the genome assemblies of two species in the *Silphium* genus and this resource was used to genotype different populations for phylogenetic studies, population structure and association mapping. The phased reference genomes were assembled based on an interspecific cross between *S. integrifolium and S. perfoliatum*, and they contain seven chromosomes each, as previously indicated by a cytological study[19]. 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/1 C and the largest plant chromosome assembled to date, reaching 14.14 Gb[20]. The authors argued that the secondary diagonal is evidence for a helical structure unique to large chromosomes in interphase

Table 2 | Loci detected in both GWAS models and multiple environments explaining more than 10% of the phenotypic variance

SNP[a]	Alleles	Trait[b]	Model	Environment[c]	N	MAF	p-value	PVE(%)	Gene	Annotation
Chr01_1583902655	T/C	RFC	BLINK	Combined	230	0.05	5.54E-09	29.39	*silint01g14554*	MATE family efflux transporter
		RFC	FarmCPU	Combined	230	0.05	3.98E-08	29.39		
		RFC	BLINK	KS24	190	0.05	1.82E-15	29.39		
		RD	FarmCPU	AL25	194	0.06	3.84E-10	24.30		
		SNC	BLINK	KS25	198	0.05	2.29E-08	25.01		
		SNC	FarmCPU	KS25	198	0.05	4.86E-08	25.01		
Chr01_695173749	G/A	RFC	BLINK	Combined	230	0.04	4.23E-10	22.82	*silint01g06242*	A/β-hydrolase
		RFC	FarmCPU	Combined	230	0.04	9.82E-14	22.82		
		SNC	FarmCPU	AL25	194	0.05	4.18E-12	23.37		
Chr02_1160355565	A/T	RD	BLINK	Combined	225	0.19	4.93E-10	11.22	*silint02g11897*	Putative transcription factor FAR
		RD	FarmCPU	Combined	225	0.19	2.62E-13	11.22		
Chr03_721440854	G/A	RFC	BLINK	AL25	196	0.17	1.61E-08	31.20	*silint03g07384*	Unknown protein
		RD	FarmCPU	KS25	199	0.16	3.00E-08	28.48		
Chr03_78530134	G/A	HI	BLINK	Index	183	0.03	2.15E-11	12.45	intergenic window	
		HI	FarmCPU	Index	183	0.03	1.93E-13	12.45		
Chr03_910576785	A/G	SNC	BLINK	Combined	225	0.06	9.05E-08	18.28	*silint03g09429*	DDB1 and CUL4 associated factor 4
		SNC	FarmCPU	Combined	225	0.06	7.14E-09	18.28		
Chr04_222292225	G/A	RFC	BLINK	KS25	206	0.03	3.73E-15	23.03	*silint04g01657*	ACT domain containing protein ACR4
		RFC	FarmCPU	KS25	206	0.03	1.15E-08	23.03		
Chr04_603756119	T/A	SNC	BLINK	KS25	198	0.03	1.24E-09	18.31	intergenic window	
		SNC	FarmCPU	KS25	198	0.03	2.75E-14	18.31		
Chr05_11691107	G/A	RFC	BLINK	Combined	230	0.02	7.52E-19	14.38	intergenic window	
		RFC	FarmCPU	Combined	230	0.02	1.72E-16	14.38		
Chr05_794020568	A/C	RFC	BLINK	AL25	196	0.03	2.77E-08	12.80	*silint05g07490* (9 kb downstream)	Thioredoxin-like protein
		RFC	FarmCPU	AL25	196	0.03	3.73E-13	12.80		
		SNC	FarmCPU	AL25	194	0.02	5.18E-24	13.89		
Chr05_831132026	G/A	RFC	BLINK	KS25	206	0.02	2.25E-15	18.35	*silint05g07840* (36 kb downstream)	Unknown protein
		RFC	FarmCPU	KS25	206	0.02	3.79E-12	18.35		
Chr06_274305900	G/A	AI	BLINK	Index	183	0.12	5.18E-10	12.75	*silint06g02767* (26 kb downstream)	Glutathione transferase
		AI	FarmCPU	Index	183	0.12	1.25E-14	12.75	*silint06g02768* (21 kb upstream)	ALP1-like protein

[a]SNP name given as the chromosome and position in base pairs.
[b]Traits measured in field-grown plants, RD: Receptacle Diameter; RFC: Ray Floret Count, SNC: Seed Number per Capitulum. Traits extracted from geospatial databases, AI: Aridity Index; HI: Heat Index.
[c]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. Markers are selected in the GWAS by combining iterative steps of Fixed Effect Model (FEM) with Random Effect Model (REM) for FarmCPU and Bayesian Information Criterion (BIC) for BLINK. Final *p*-values are calculated using single-marker least-squares regression.
SNP Single Nucleotide Polymorphism, *Chr* Chromosome, *MAF* Minor Allele Frequency, *N* Number of samples, *PVE* Phenotypic variance explained.

nuclei, which exists to prevent DNA entanglement and promote efficiency for transcriptional regulation. The fact that we observed similar structure independently in another species with large chromosomes, provides additional evidence of this phenomenon. The structure seems to be 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 at additional 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*[20]. 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 among *Silphium*, we reconstructed the genus phylogeny with 14 species, including multiple samples in each taxon. Overall, the phylogeny had great accordance to the first *Silphium* phylogeny based on ITS and ETS sequencing data[21], 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[21] and our study was a lack of resolution in the *Silphium* section. *S. integrifolium*, *S. wasiotense* and *S. perfoliatum* were previously grouped together[21] and *S. radula, S. asteriscus* and *S. trifoliatum* formed another group[21]. 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[8]. 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[8]. Further development of taxonomic and evolutionary knowledge of *Silphium* must be prioritized to guide

allele mining and hybridization efforts of *Silphium* breeders 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, 156 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 they 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 *S. integrifolium var. neglectum* across Illinois, Wisconsin and Indiana, indicating a possible differentiation of this variety from the other populations[22]. When assuming $K = 4$, one individual in Texas appears to represent a distinct southwest population, which matches the groups identified previously[23]: 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[23], which would explain the merger 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 adaptive alleles[24], potentially accelerating breeding for traits such as environmental adaptation.

Four major loci associated with ray floret count, number of seeds per capitulum and receptacle diameter were detected in this study. The capitulum in most Asteraceae is a compound inflorescence with many small flowers (florets) borne on a single disk-like receptacle. The ray floret count is a particularly important target for selection because differently from other genera in Asteraceae, in *Silphium* only the ray florets are female-fertile and produce seeds, while the more numerous disk florets produce only pollen[25] (Supplementary Fig 15). Therefore, one of the major goals in *Silphium* domestication is to increase both the proportion of female florets ("feminization") and the total number of florets per head, which can potentially increase seed yields[26]. Targeting the four loci identified here for selection, should help accomplish this goal and accelerate the domestication of this nascent crop.

The four loci associated with capitulum traits (seed number per capitulum, ray floret count and receptacle diameter) are in four genes: *silint01g06242* (α/β-hydrolase), *silint01g14554* (MATE family efflux transporter), *silint03g07384* (unknown protein) and *silint04g01657* (ACR4-related protein). α/β-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 α/β-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[27]. The multidrug and toxic compound extrusion (MATE) transporters is 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[28]. 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[29].

The SNP Chr03_78530134 was identified for heat index and explained 12.45% of the variance. This SNP is located 76,690 bp upstream from the gene *silint03g00999*, a beta-N-acetylhexosaminidase that has been shown to be involved in fruit softening, ripening, and response to pathogens[30]. The SNP Chr06_274305900 was detected in association with the aridity index, explaining 12.75% of the variance. This variant is located 25,740 bp downstream the gene *silint06g02767* and 20,878 bp upstream the *silint06g02768*. These two genes had no functional annotations assigned with InterproScan, however a blastp search at NCBI indicated that they have similarity to *Helianthus annuus* glutathione transferase [percent identity: 37%] and ALP1-like protein [percent identity: 81%], respectively. Glutathione transferases are expressed in several stages of plant development and have increased expression under stress conditions, such as drought[31]. The ALP1 gene is derived from Harbinger transposases and has become involved in several roles in genome regulation. In *Arabidopsis*, mutations in this gene lead to down-regulation of other genes involved in disease resistance and stress response[32]. Despite the knowledge of some functions of the candidate genes identified here, their potential role and the effect of the variants on the abiotic stress tolerance and capitulum morphology in *Silphium* is not clear yet. Therefore, further functional studies are needed to expand our knowledge and validate these candidates.

In attempting to domesticate a plant species 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 has been reported in soybean[33]. An important question for those trying to create new crops is: can we obtain significant improvement in agronomic traits without losing environmental adaptation alleles? With the resources now available, crop domestication can take a 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 genotypes that preserve both domestication alleles (such as the candidate *Silphium* capitulum loci described above) and abiotic stress related alleles (such as candidate *Silphium* aridity and heat index 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. In fact, when performing a linkage disequilibrium analysis between capitulum and aridity/ heat index loci, there was no evidence that the capitulum loci and the environmental index loci are co-segregating, indicating that changes in one will not affect the other, which opens the possibility of combining favorable alleles for both groups of traits (Supplementary Discussion 1, Supplementary Fig 16).

Aiming to understand the implications of breeding history on the ability to detect significant loci, we performed a sensitivity analysis on subsets of the SAP. In the genome scans with *S. integrifolium* genets only, three of the four capitulum loci were still detected, but for the wild or breeding genets only, none of the loci were significant. This is likely a direct impact of the reduction in the favorable allele frequency on the wild subset and the drastic reduction in sample size in the breeding genets, which are determinant factors for the power to detect significant loci[34]. This also shows that our ability to detect the capitulum loci in the SAP is a result of the breeding history of several genets, which increased the overall frequency of favorable alleles. However, high allele frequency alone is not enough for detection, as observed by the lack of significance in the breeding only subgroup. The increase in allele frequency needs to be accompanied by a larger sample size for loci to be correctly detected.

This result challenged our initial view that improvements in genome assembly and genotyping offer a way to directly use gene discovery to leapfrog phenotypic selection in the de novo domestication of wild species. However, we would not have been able to identify loci for capitulum domestication using only the subset of 206 geographically diverse wild genets. Given the very low allele frequency for

the favorable allele in wild populations, a much larger GWAS panel would have been required. Therefore, we hypothesize that a more efficient strategy for genomics-enabled domestication of a completely wild species would be to first perform a few cycles of phenotypic selection on the traits of interest in a wild collection to increase the frequency of potential rare alleles and increase the chances of detecting associations. Cycles of selection could be done while the candidate crop's genome is being sequenced and assembled and should yield insights about basic horticulture and breeding techniques.

The SAP contains 25 individuals classified as *S. radula*, 3 *S. perfoliatum* and 5 *S. integrifolium* × *perfoliatum*. One of the reasons they were included in the panel is that we wanted to maximize the potential diversity in the panel and include closely related species that are used in the domestication program. Additionally, *Silphium* species are known to naturally hybridize, and this mixture has complicated the placement of species and subspecies boundaries[35]. The full population had a significant higher power to detect loci than the subsets, which is consistent with other observations that statistical power improves in populations composed of mixed species and bigger populations[34]. The fact that we were able to detect only three of the four capitulum loci in the *S. integrifolium* subset and none of the four in the wild subset, shows that genomic scans in purely wild or purely breeding genets could miss important loci.

Therefore, combining the different groups into a single diverse population can be more effective for multi-trait scans and provide a better overview of important loci early in the domestication program. In parallel with recommending cycles of phenotypic selection preceding genome association studies for future de novo domestication projects, we recommend early exploratory interspecific hybridization of the candidate species with congeners. This would help create a multi-species/interspecific association panel with heightened phenotypic variance involving alleles that could be brought into the breeding population and prepare for long term introgression breeding.

The assembly of reference genomes for *S. integrifolium* Michx. and *S. perfoliatum* L. has been essential for the development of the genotyping strategy used in the populations studied here. The functionality of the targeted sequencing approach was confirmed by using it to advance the phylogenetic characterization of the *Silphium* genus and to discover environmental and phenotypic associations with markers and candidate genes. These results give us confidence that both marker-assisted breeding and genomic selection can soon be implemented by *Silphium* breeders. Genomic-assisted breeding methods have proven efficacy in accelerating the domestication of other perennial crops, such as intermediate wheatgrass[36], and these strategies can be applied to other wild species, orphan crops with limited resources and major crops with large and complex genomes that need more cost-effective solutions.

One question that may arise from the features of the *Silphium* genomes 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[37]. It should be a priority to survey the recombination landscape for *Silphium* chromosomes for this pattern, because if present, large, stable pericentromeric regions are likely to have accumulated genetic load like other recombination coldspots[38] and could have important impact on *Silphium* yield and domestication traits. In *Arabidopsis* and maize, it has been demonstrated that heterozygous regions surrounded by a homozygous background attract more crossovers[39]. If the same is true for *Silphium*, with the availability of genomes and molecular markers, we may have the ability to make strategic crosses to increase the likelihood of crossing over events within low-recombination regions. Future efforts to screen *Silphium* germplasm collections and breeding populations for rare pericentromeric recombinants would build on our investment in basic genomics and genotyping. We suggest that genomic investments early in any crop domestication will make it possible to discover and maintain types of genetic diversity that could otherwise be difficult to re-introduce at a later stage of domestication.

Methods

Plant material

The genets 'Bad Astra' (*S. integrifolium* – HAP1) and '1JHW' (*S. perfoliatum* – HAP2) were crossed to generate an F_1 interspecific hybrid '3ZZP' for a haplotype-phased assembly in a trio binning approach[40] (Supplementary Figs. 17 and 18). '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 (Supplementary Methods 1).

To expand our knowledge about the evolution and species relationships in *Silphium*, 41 samples from 14 different *Silphium* species, and eight outgroup samples from three closely related members of the Asteraceae family (*Lindheimera texana, Engelmannia peristenia and Helianthus maximiliani*) were obtained from collaborators or commercial nurseries (Supplementary Data 4).

To investigate the genetic architecture of domestication and adaptation traits, the *Silphium* association panel (SAP) was developed at TLI by identifying 258 genets from a larger long-term ex situ collection (Supplementary Fig. 17, Supplementary Data 5). The SAP represents the collection's geographic diversity and the full range of phenotypic variation. To avoid confusion in nomenclature, here we used the term "genet" to refer to each accession in the panel, since it is a term used to define a genetically unique organism[41]. We refer to "haplotype" as the entire set of a haploid genome (n) inherited from one of the parental lines and used "genotype" to denote a specific combination of alleles.

The SAP is composed of 225 *S. integrifolium*, 25 *S. radula*, 3 *S. perfoliatum* genets and 5 *S. integrifolium* × *perfoliatum* hybrid genets. In the domestication program at TLI, the main interest is on *S. integrifolium*, but *S. perfoliatum* and its hybrids were included because *S. perfoliatum* was used to develop the hybrid used in genome assembly (described above) and as a source of disease resistance genes via introgression breeding. *S. radula* genets were included because they had originally been identified as *S. integrifolium* but were later re-identified as *S. radula*. 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 with clones of 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 "Genome-wide association study").

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 long read sequencing and nuclei isolation for 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 the Omni-C library preparation following the Dovetail Omni-C kit protocol (Dovetail Genomics, California, USA). Leaf tissue from the parents 'Bad Astra' and '1JHW' was collected and DNA was extracted using the modified CTAB extraction[42] for Illumina short read sequencing. For the *Silphium* Association Panel and the phylogenetic study samples, 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 3). 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 flash frozen in liquid nitrogen and ground using a Freezer Mill, and the RNA was extracted with CTAB[43]. The Zymo Clean and Concentrator kit was used for RNA cleanup after extraction (Zymo Research, California, USA).

Genome sequencing and assembly

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

Initial estimation of the nuclear genome size was performed with jellyfish (v2.3.10)[44] to count k-mers and visualization with GenomeScope (v2.0)[45] and Smudgeplot (v0.4)[45]. The short-read sequences from the parental genotypes were used to phase the two haplotypes using HiFiasm (v0.19.8)[46] following the trio binning method[40]. Dovetail Omni-C reads were separately mapped to the HAP1 and HAP2 contig sets using bwa-mem (v0.7.17)[47] and chromosome scale scaffolding was performed with Juicer (v1.6)[48]. Information about the fraction of paired end reads that are found at different distances from one another in 1 kb windows were obtained from the mapped Omni-C reads using the HOMER[49].

The contigs were then oriented, ordered, and joined together into chromosomes per haplotype using the Omni-C data. Chromosome joints were padded with 10,000 'N' calls and the assemblies were subsequently polished using RACON[50].

Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and oriented to make sure that they were properly positioned in the 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[51]. The LTR assembly index (LAI) was used to evaluate the continuity and completeness of the genome assemblies[52]. The genome structure and synteny were visualized using the R package Genespace (v1.3.1)[53].

Repeat and gene annotation

Transposable elements prediction and annotation was performed with the Extensive de novo TE Annotator (EDTA) pipeline (v1.9.3)[54] using the following parameters: "--overwrite 0 --sensitive 0 --anno 1 --evaluate 0 --threads 25".

Illumina TruSeq stranded RNAseq libraries constructed for the F1 Hybrid '3ZZP' were sequenced at HudsonAlpha with an Illumina NovaSeq6000. The RNA sequencing reads from the eleven different tissue types were used for gene prediction in BRAKER3[55]. 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 Phytozome (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 (v5.66-98.0)[56].

Chloroplast assembly and annotation

The chloroplasts from *S. integrifolium* and *S. perfoliatum* were assembled with the OatK (v1.0) pipeline (https://github.com/c-zhou/oatk) utilizing the angiosperm database. The chloroplast genome was then polished using RACON[50]. The assembled genomes were annotated using GeSeq (v2.0.3)[57] and visualized using OGDRAW (v1.3.1)[58], both tools available at GeSeq (https://chlorobox.mpimp-golm.mpg.de/geseq.html).

Comparative genomics

To contextualize the *Silphium* genomes within the Asteraceae family, 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.v1 and HA412-HO.v2 assemblies[59,60] and the *Scaevola taccata* assembly ST1.0[61]. 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) was used for the gene annotation.

SAP sequencing and variant calling

The *Silphium* association panel was sequenced with two combined strategies. The first was low-coverage sequencing aiming for an average of 1 × genome coverage. Libraries were prepared with the Twist 96-Plex kit (Twist Bioscience, California, USA) and sequencing was performed on the Illumina NovaSeq 6000. The second strategy was targeted sequencing aiming to enrich sequencing in regions containing SNPs or genes. In total, 50,000 probes were designed, then synthesized by Twist Bioscience and used in the Twist target enrichment standard hybridization (Supplementary Methods 2; Supplementary Data 6). For the enrichment, sequencing libraries were prepared with Twist Flex Prep kit and after hybridization, the samples were sequenced on an Illumina Novaseq 6000 (Supplementary Fig. 19).

The two datasets, low coverage and targeted sequencing, were combined for each individual and checked for total number of sequencing reads, mean Phred scores and GC content using MultiQC[62]. Eight samples with less than 13 million reads and abnormal GC content distribution were removed from the analysis, and the resulting panel of 258 individuals was then processed for variant calling. The calling was based on the Khufu platform to generate a hapmap[63]. Khufu integrates several tools, including read mapping with bwa (v0.7.17)[47], variant calling and quality control with BCFtools (v1.19)[64] and imputation with STITCH (v1.8.4)[65]. 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 Fig. 20).

Reconstructing the phylogeny of *Silphium*

To reconstruct the genus phylogeny, 49 additional samples (41 *Silphium* and eight outgroup samples) were sequenced using the targeted sequencing method. In addition, six *S. radula*, three *S. perfoliatum*, six *S. integrifolium* and five *S. integrifolium* × *perfoliatum* from the SAP were included. Data from the two sunflower genomes HA412HO and XRQ and *Scaevola taccada* was also included, resulting in 72 samples in total for this analysis.

To select the loci from the 50,000 targeted sequencing panel, 7000 loci with reasonable depth (15X <depth <150X) were selected based on SAMtools depth (v1.19)[64]. A target file was created by extracting a sequence of 500 bp from the 'Bad Astra' assembly for each of the 7000 loci using the function getfasta from BEDtools (v2.28.0)[66]. To assemble the target regions, the fastq data was mapped to the targets fasta sequences using HybPiper (v2.3.2)[67]. To identify paralogs, the command HybPiper paralog_retriever was used, and the paralogs were removed from the panel. In addition, only the loci 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 Data 7 and 8). To align the sequences, MAFFT (v7.525)[68] was used with the following parameters: --preservecase --maxiterate 1000. Trimal (v1.5.0)[69] was subsequently used to remove poorly aligned sections from the aligned sequences with -gt .5. To build the gene trees, IQTree (v2.3.6)[70] was used with the ModelFinder Plus option[71] and 1000 bootstraps. To collapse the gene tree with branches below 50% bootstrap value, the nw_ed tool from Newick Utilities (v1.6)[72] was used with the parameter $b < 50$. To build the species tree from the gene trees, ASTRAL (v5.7.8)[73] was used with default parameters. Tree files were visualized using FigTree (v1.4.4)[74].

Kinship, population structure and ancestry

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 (v5.0)[75] and used to analyze the kinship and population structure among the accessions. PCA results were plotted with the R package ggplot2 (v3.5.2)[76] and the hierarchical cluster based on the kinship matrix was plotted with the R package ggtree (v3.12.0)[77]. The spatial ancestry analysis for *S. integrifolium* was conducted 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 (v1.1.0)[78], 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.

Genome-environment association study(GEA)

GEA scans on two environmental indexes, heat index (HI) and aridity index (AI) were performed to identify genomic regions associated with environmental variables. For HI, mean temperatures were obtained for the 183 georeferenced genets from Worldclim (https://www.worldclim.org/) at a resolution of 30 arcsec (1 km^2) considering the historical averages from 1970 to 2000 (WorldClim version 2.1, Supplementary Fig. 21). Data extraction and preparation were performed using the R packages geodata (v0.6-2), terra (v1.8-42) and raster (v3.6-32)[79–81]. The HI was calculated using the evapotranspiration model[82,83], 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 (Eq. 1).

$$HI = \sum_{i=1}^{k} \frac{Tm^{1.514}}{5} \tag{1}$$

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)[84] 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 arcsec (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.

The association between the indexes and SNPs was performed using the R package GAPIT(v3.4.0)[85], using two different statistical models (FARMCPU and BLINK). Principal components and kinship were accounted as cofactors in the models. 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.

Genome-wide association study (GWAS)

A GWAS analysis was performed with the 449554 SNPs for four phenotypes in the *Silphium* association panel field trials: 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 4, Supplementary Fig. 22). Best linear unbiased estimation (BLUE) was calculated in single and the combination of environments (AL24, KS24, AL25, KS25, Combined) using the R package lme4 (v1.1-37)[86] with the following model:

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

y_{ijkl} corresponds to the response value for the ith genet in the jth block within kth replication within lth environment. μ, g_i, $b_{j(k)}$, r_k, e_l and ge_{il} represent the effect of the mean, the genet (fixed), the incomplete block (random), replicate (random), the environment (random) and the interaction genet × 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[87]:

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

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 and significance threshold was defined by a Bonferroni correction. To understand the effect of population subgroups on the associations, a sensitivity analysis was performed by rerunning the genome scans in subsets of the SAP: *S. integrifolium* individuals only ($N = 236$), wild individuals only ($N = 206$), and breeding individuals only ($N = 47$).

Linkage disequilibrium analysis was performed for the major loci identified with the R packages GAPIT3(v3.4.0) and SNPRelate (v3.22)[88]. All analyses conducted with R packages were performed in the R version 4.4.2[89]. 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. To understand the effect of the alleles at each locus on the phenotype, the SNP genotypes were retrieved and tested as predictors for the phenotype in a linear regression and compared with a with a two-tailed Student's *t*-test.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All sequencing data generated in this project has been deposited at the National Center for Biotechnology Information. Raw sequencing data for genome assembly is under Bioproject PRJNA1417267. Sequencing data for the *Silphium* association panel and phylogeny is under Bioproject PRJNA1426741 and PRJNA1426744, respectively. Final genome assembly for *S. perfoliatum* (1JHW) is under PRJNA1422233 and for *S. integrifolium* ('Bad Astra') is under PRJNA1422234. Genome annotations, chloroplast assembly, and annotation are available at Figshare [https://figshare.com/s/f75787ae1adf175ad955]. Source data are provided with this paper.

Code availability

Codes to generate the results in this research are deposited at Github [https://github.com/renanso/silphium_genome_codes][90]. The R package for probe design for targeted sequencing is also available at Github [https://github.com/renanso/enrich/][91].

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. USA* **104**, 8641–8648 (2007).
2. 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).
3. Glover, J. D. et al. Agriculture. Increased food and ecosystem security via perennial grains. *Science* **328**, 1638–1639 (2010).
4. Bayer, P. E. et al. Sequencing the USDA core soybean collection reveals gene loss during domestication and breeding. *Plant Genome* **15**, e20109 (2022).
5. Zsögön, A. et al. De novo domestication of wild tomato using genome editing. *Nat. Biotechnol.* **36**, 1211–1216 (2018).
6. Yu, H. et al. A route to de novo domestication of wild allotetraploid rice. *Cell* **184**, 1156–1170.e14 (2021).
7. Yu, H. & Li, J. Breeding future crops to feed the world through de novo domestication. *Nat. Commun.* **13**, 1171 (2022).
8. Brock, M. & Weakley, A. Silphium Linnaeus 1753 (Rosinweed). in *Flora of the Southeastern United States* (ed. Weakley, A. S.) (University of North Carolina Herbarium, 2020).
9. 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).
10. 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).
11. 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).
12. Reinert, S., Van Tassel, D. L., Schlautman, B., Kane, N. C. & Hulke, B. S. Assessment of the biogeographical variation of seed size and seed oil traits in wild *Silphium integrifolium* Michx. genotypes. *Plant Genet. Resour.* **17**, 427–436 (2019).
13. 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. *Am. Oil Chem. Soc.* **101**, 663–674 (2024).
14. 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**, 1829160 (2018).
15. 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* **2024**, 8464635 (2024).
16. Weaver, J. E. Classification of root systems of forbs of grassland and a consideration of their significance. *Ecology* **39**, 393–401 (1958).
17. 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).
18. Bohra, A. et al. Reap the crop wild relatives for breeding future crops. *Trends Biotechnol.* **40**, 412–431 (2022).
19. Fisher, T. R. & Cruden, R. Chromosome number and observations in the genus *Silphium*. *Ohio J. Sci.* **62**, 258–259 (1962).
20. 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).
21. 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).
22. Settle, R. & Fisher, W. J. The varieties of *Silphium integrifolium*. *Rhodora* **72**, 536–543 (1970).
23. 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).
24. Norris, E. T. et al. Admixture-enabled selection for rapid adaptive evolution in the Americas. *Genome Biol.* **21**, 29 (2020).
25. 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).
26. 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).
27. 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).
28. 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).
29. 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).
30. Irfan, M. et al. N-glycan remodeling by α-D-mannosidase and β-D-N-acetylhexosaminidase regulates fruit softening, redox balance, and post-harvest pathogen resistance. *Plant J.* **124**, e70629 (2025).
31. Xu, J. et al. Transgenic Arabidopsis plants expressing tomato glutathione S-transferase showed enhanced resistance to salt and drought stress. *PLoS ONE* **10**, e0136960 (2015).
32. Liang, S. C. et al. Kicking against the PRCs - A domesticated transposase antagonises silencing mediated by Polycomb group proteins and is an accessory component of Polycomb repressive complex 2. *PLoS Genet.* **11**, e1005660 (2015).
33. Hyten, D. L. et al. Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl. Acad. Sci. Usa.* **103**, 16666–16671 (2006).

34. Hamazaki, K. et al. Choosing the optimal population for a genome-wide association study: A simulation of whole-genome sequences from rice. *Plant Genome* **13**, e20005 (2020).
35. Clevinger, J. A. New Combinations in *Silphium* (Asteraceae: Heliantheae). *Novon* **14**, 275–277 (2004).
36. 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).
37. Brazier, T. & Glémin, S. Diversity and determinants of recombination landscapes in flowering plants. *PLoS Genet.* **18**, e1010141 (2022).
38. Rodgers-Melnick, E. et al. Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proc. Natl. Acad. Sci. USA* **112**, 3823–3828 (2015).
39. Mikhailov, M. E., Boideau, F., Szymanska-Lejman, M., Botnari, V. & Ziolkowski, P. A. Enhancing local meiotic crossovers in Arabidopsis and maize through juxtaposition of heterozygous and homozygous regions. *Nat. Plants* **11**, 1769–1784 (2025).
40. Koren, S. et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **36**, 1174–1182 (2018).
41. Zhang, X. et al. Establishment and optimization of genomic selection to accelerate the domestication and improvement of intermediate Wheatgrass. *Plant Genome* **9**, plantgenome2015.07.0059 (2016).
42. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* **19**, 11–15 (1987).
43. 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).
44. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
45. 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).
46. 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).
47. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
48. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C Experiments. *Cell Syst.* **3**, 95–98 (2016).
49. 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).
50. 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).
51. 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).
52. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
53. Lovell, J. T. et al. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *Elife* **11**, e78526 (2022).
54. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
55. Gabriel, L. et al. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. *Genome Res.* **34**, 769–777 (2024).
56. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
57. Tillich, M. et al. GeSeq - versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* **45**, W6–W11 (2017).
58. 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).
59. Badouin, H. et al. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* **546**, 148–152 (2017).
60. Hübner, S. et al. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nat. Plants* **5**, 54–62 (2019).
61. Shen, F. et al. Comparative genomics reveals a unique nitrogen-carbon balance system in Asteraceae. *Nat. Commun.* **14**, 4334 (2023).
62. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
63. Korani, W. et al. De novo QTL-seq Identifies Loci Linked to Blanchability in Peanut (Arachis hypogaea) and Refines Previously Identified QTL with Low Coverage Sequence. *Agronomy* **11**, 2201 (2021).
64. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
65. Davies, R. W., Flint, J., Myers, S. & Mott, R. Rapid genotype imputation from sequence without reference panels. *Nat. Genet.* **48**, 965–969 (2016).
66. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
67. Johnson, M. G. et al. HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Appl. Plant Sci.* **4**, 1600016 (2016).
68. Katoh, K., Misawa, K., Kuma, K.-I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
69. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
70. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
71. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
72. Junier, T. & Zdobnov, E. M. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics* **26**, 1669–1670 (2010).
73. Mirarab, S. et al. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541–i548 (2014).
74. Rambaut, A. *FigTree. A Graphical Viewer of Phylogenetic Trees*. https://tree.bio.ed.ac.uk/software/figtree/ (2018).
75. Bradbury, P. J. et al. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
76. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer, 2009).
77. 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).
78. 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).

79. Hijmans, R. J., Barbosa, M., Ghosh, A. & Mandel, A. Geodata: Access Geographic Data. R package version 0.6-2. https://github.com/rspatial/geodata (2026).
80. Hijmans, R. J. et al. Terra: spatial data analysis. *R. package version* **1**, 8–42 (2025).
81. Hijmans, R. J., van Etten, J., Sumner, M., Cheng, J. & Baston, D. Raster. *R. package version* **3**, 6–32 (2025).
82. Thornthwaite, C. W. An approach toward a rational classification of climate. *Geogr. Rev.* **38**, 55–94 (1948).
83. 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).
84. Zomer, R. J., Xu, J. & Trabucco, A. Version 3 of the Global Aridity Index and Potential Evapotranspiration Database. *Sci. Data* **9**, 409 (2022).
85. Lipka, A. E. et al. GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28**, 2397–2399 (2012).
86. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
87. 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).
88. Zheng, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
89. R. Core Team. *R: A Language and Environment for Statistical Computing*. *R Foundation for Statistical Computing*. (2021).
90. Souza, R. S. et al. Assembly of *Silphium* interspecific hybrid genomes opens the genus to phylogenomics, ecogenomics, and molecular breeding, *github.com/renanso/silphium_genome_codes*, https://doi.org/10.5281/zenodo.20562548 (2026).
91. Souza, R. S. et al. Assembly of *Silphium* interspecific hybrid genomes opens the genus to phylogenomics, ecogenomics, and molecular breeding, *github.com/renanso/enrich*, https://doi.org/10.5281/zenodo.20562561, (2026).

Acknowledgements

The authors thank Patricia Sanmartin, Zachary Myers, Haley Hale and Zachary Meharg at HudsonAlpha Institute for Biotechnology for assistance in DNA extraction, library preparation, and sequencing data generation. 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 low coverage sequencing and targeted 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., and S.W. contributed to manuscript preparation and revision. All the authors read and approved the final manuscript.

Funding

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

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-75205-3.

Correspondence and requests for materials should be addressed to David L. Van Tassel.

Peer review information *Nature Communications* thanks Wei Fan, Maddie James and Arun Sethuraman for their contribution to the peer review of this work. A peer review file is available.