

Phylogenomic investigation of safflower (*Carthamus tinctorius*) and related species using genotyping-by-sequencing (GBS)

Somayeh Sardouei-Nasab (✉ sardoueinasabs@gmail.com)

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)

Zahra Nemati

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)

Ghasem Mohammadi-Nejad

Shahid-Bahonar University of Kerman

Reza Haghi

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)

Frank R. Blattner

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)

Article

Keywords: Domestication, Ecotypes, Genotyping-by-sequencing (GBS), Phylogeny, Safflower, Wild progenitor

Posted Date: June 29th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1745176/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on April 17th, 2023. See the published version at <https://doi.org/10.1038/s41598-023-33347-0>.

Abstract

Safflower (*Carthamus tinctorius*) is a source of high-quality edible oil growing in moisture-limited environments. Despite its economic importance, the relationships to close wild species in *Carthamus* and the presence and relationships of ecotypes within safflower are still not fully clarified. Here we use genotyping-by-sequencing to identify the wild progenitor of *C. tinctorius*, infer phylogenetic relationship within series *Carthamus* and identify groups of closely related lineages within cultivated safflower. Phylogenetic and population genetic analyses found *C. palaestinus* to be the closest relative and single progenitor of *C. tinctorius*, which confirms the Levant as the area of domestication of the crop. Flow cytometry showed all analyzed samples of *C. oxyacantha*, *C. palaestinus* and *C. tinctorius* to be diploid ($2n = 2x = 24$) with 2C genome sizes of 2.4 to 2.7 pg. Analyses of a set of 114 worldwide distributed safflower accessions arrived at two to five genetic groups, which showed, however, no correlation with the geographic origins of these accessions. From this, we conclude that the trade of safflower seeds resulted in multiple introductions of genotypes from the Levant to other areas with suitable climate conditions for the plant, as well as exchange of genotypes among these areas.

Introduction

Safflower (*Carthamus tinctorius* L., Asteraceae) is used to produce an orange dye that is sometimes used as a substitute for saffron and as an oilseed crop that grows in hot and dry climates tolerating saline soils [1]. Safflower was domesticated in the Fertile Crescent over 4,000 years ago [1,2], most probably for its use as colorant, but is nowadays cultivated in suitable areas of Asia, northeastern Africa and Europe [3,4] with a total cultivation area of 1.1 million hectares resulting in seed yield of approximately 1 million tons per year [5]. Safflower's high-quality edible oil contains higher amounts of unsaturated fatty acids like oleic and linoleic acids than other seed oils. Despite the growing commercial importance of safflower, genetic diversity and the parental origin of this species are not fully clarified mainly due to the low variability of up-to-now applied markers. Among the species of *Carthamus*, *C. oxyacantha* and *C. palaestinus* – both with $2n = 24$ chromosomes like *C. tinctorius* – have been reported as the most likely wild ancestors of safflower [3,6-9]. Wild relatives together with landraces carry valuable characteristics for crop breeding, such as biotic and abiotic stress resistances [10]. Thus, hybridization of safflower with its closest relatives could be used for the introgression of traits into the crop and can play a significant role in future breeding of cultivars. Some species of *Carthamus* including *C. lanatus*, *C. palaestinus* and *C. oxyacantha* have been reported to have the potential to cross with safflower and produce fertile offspring [6,7,11]. Due to partly low reproducibility of applied markers and conflict between published data sets [3,9,12-17], we here use genome-wide single-nucleotide polymorphisms (SNP) obtained through genotyping-by-sequencing [18] (GBS) on a diverse set of *C. tinctorius* accessions and six wild *Carthamus* species for a high-resolution phylogeny. The objectives of our study are (i) to infer phylogenetic relationships of *Carthamus* species closely related to safflower, i.e. section *Carthamus*, (ii) to identify the closest wild relative(s) of safflower to enable breeding for the increase of genetic diversity regarding fatty acid content, drought tolerance and other agro-morphological traits, (iii) to look into

genetic diversity of safflower accessions, and (iv) to see if the three to ten postulated morphology and geography-defined groups [9,19-24] are genetically distinct and can be recovered using genomic data.

Results

Phylogenetic inference

In an initial phylogenetic analysis, we included six *Carthamus* species, i.e. *C. boissieri*, *C. glaucus*, *C. lanatus* (with three subspecies), *C. oxyacantha*, *C. palaestinus* and *C. tenuis* together with a small but geographically diverse subset of *C. tinctorius* to infer the closest relative of safflower. The GBS-derived data matrix consisted of 108,985 aligned characters of which 3468 were parsimony informative. We used neighbor-joining (NJ), Bayesian phylogenetic inference (BI) and maximum parsimony (MP) analyses which resulted in identical (BI, MP) or compatible (BI, MP, NJ) tree topologies. The MP analysis retained 2484 equally parsimonious trees with a consistency index (CI) of 0.68 and a retention index (RI) of 0.95. The strict consensus of these trees, summarizing MP (bootstrap, bs) and BI (posterior probabilities, pp) support values, is provided in Fig. 1.

This unrooted tree consists of two major clades, one including the 24-chromosome species (chromosome numbers according to Vilatersana et al. [25]) *C. oxyacantha*, *C. palaestinus*, and *C. tinctorius* together with $2n = 20$ *C. glaucus*, while clade two harbors $2n = 20$ *C. tenuis* and *C. boissieri*, together with the $2n = 44$ allotetraploid *C. lanatus*. *Carthamus tinctorius* groups within *C. palaestinus*, indicating the latter to be the closest relative of safflower and probably the wild progenitor of the crop. The two Iranian *C. oxyacantha* accessions from Isfahan and Kashan are separated from the other members of this species and fall within the *C. palaestinus*/*C. tinctorius* clade, indicating introgression from these species. Support values for the branches separating the species are generally very high throughout the tree (Fig. 1).

Since, *C. tinctorius* was found in a clade with *C. oxyacantha*, *C. palaestinus* and *C. glaucus*, we included these wild species in our analysis of 114 safflower accessions. The alignment for this dataset consisted of 435,024 characters, which resulted in four MP trees (CI = 0.30, RI = 0.51) that were rooted with the six non-introgressed *C. oxyacantha* accessions. In this tree (Fig. 2) *C. palaestinus* groups within *C. tinctorius* with 100% bootstrap support for the clade unifying these species. Within *C. tinctorius* bootstrap support values >75% occur only for a few and rather small clades of safflower accessions. Although *C. tinctorius* accession formed five larger groups (Fig. 2), no clear geographic pattern was found, i.e. in most clades, safflower accessions from diverse geographical regions were united.

Population assignment analysis and Principal Component Analysis (PCA)

In addition to tree-based estimations of relationships of *C. tinctorius* we also used Bayesian population assignment and principal component analyses. *Carthamus oxyacantha*, *C. palaestinus*, *C. glaucus* and *C. tinctorius* individuals were subjected to Structure-like analyses in Lea. Evanno's Δk suggested $K = 3$ as the best fitting category number that was used for the final analysis (Fig. 3). The analysis confirmed our tree-based assumption that two of the *C. oxyacantha* accessions (from

Isfahan and Kashan, Iran) were introgressed by genotypes from the *C. palaestinus*/*C. tinctorius* group and in addition identified also a third accession (from Arak, Iran) showing traces of gene flow. *Carthamus palaestinus* (from Israel) has the greatest similarity with *C. tinctorius*. *Carthamus glaucus* (from Lebanon) is clearly distinct from the former two species although some allelic overlap occurs with one Russian and two Iranian safflower accessions (Fig. 3) and two individuals of *C. glaucus* share alleles with *C. palaestinus*/*C. tinctorius*.

Also PCA resulted in three groups of populations (Fig. 4). The first two principal components explained 42.8% of the total molecular variation. PCA results were similar to those of the phylogenetic trees and Lea plots as they grouped *C. tinctorius* and *C. palaestinus* together while *C. oxyacantha* and *C. glaucus* were clearly separated from them along the first and second axis, respectively.

Analysis of genome size

Flow cytometry was used to measure genome sizes and infer ploidy levels of *C. tinctorius*, *C. oxyacantha*, and *C. palaestinus* individuals (Table 1). We found that all measured samples of *C. oxyacantha*, *C. palaestinus* and *C. tinctorius* represent diploid ($2n = 2x = 24$) 2C genome sizes ranging from 2.4 pg to 2.7 pg.

Discussion

Inferring phylogenetic relationships among closely related species is a challenge because of inter-locus phylogenetic discordance and/or the difficulty of obtaining variable markers. Despite various phylogenetic studies carried out in *Carthamus*, phylogenetic relationships within *Carthamus* are poorly understood and the closest wild relative of cultivated safflower remained elusive for a long time [26] due to limited taxon sampling and/or low variability of the applied markers [12,27-29]. Still, studies published during the last decades mostly identified *C. palaestinus* from the southwestern Fertile Crescent as the possible wild ancestor [3,8,12], which fits geographically with the earliest archeological remains of safflower that were identified in Syria [2]. Also, intra-species relationships within safflower are not well understood. Based on morphological traits like flower color, inflorescence size, plant height and leaf shape, and their geographical distribution, between five and ten centers of diversity were postulated for safflower, all harboring region-specific ecotypes [1,9,19,20,26]. Here we performed phylogenetic and population genetic analyses based on genome-wide distributed SNPs that were obtained through GBS to clarify the phylogenetic relationships of the relatives of *C. tinctorius* and to see if described ecotypes of safflower are genetically distinct. The GBS approach is known to have the potential for better resolved relationships within and among closely related species in comparison to earlier fragment size analyses or sequence comparisons of single nuclear or chloroplast DNA regions [31-34].

Our phylogenetic analyses of GBS data representing seven *Carthamus* species resulted in a tree (Fig. 1) where *C. boissieri*, *C. lanatus* and *C. tenuis* form one clade that is sister to the clade of *C. glaucus*, *C. oxyacantha*, *C. palaestinus* and *C. tinctorius*. Support values for the clades in the tree are generally very high and *C. boissieri*, *C. lanatus*, *C. oxyacantha*, and *C. tenuis* result as monophyletic species. In contrast,

C. tinctorius groups within *C. palaestinus*, and both together within a paraphyletic *C. glaucus*. Within the *C. palaestinus*/*C. tinctorius* clade also two *C. oxyacantha* accessions group, which we assumed to be hybrids between *C. oxyacantha* and *C. palaestinus* or *C. tinctorius*, a result that was confirmed by Structure-like analysis (Fig. 3). Also in *C. glaucus* two individuals show allelic similarities with *C. tinctorius* (Fig. 3), which could indicate gene flow and explains that two clades of a paraphyletic *C. glaucus* were recovered in the phylogenetic trees of which the introgressed individuals are sister to *C. palaestinus*/*C. tinctorius*. Our data indicate that hybridization between the species closely related with *C. tinctorius* is still possible on the homoploid level, as the *C. oxyacantha* individual from Isfahan (Iran) was measured with the same genome size as other individuals from this species (Table 1). The two major clades we obtained in our phylogenetic tree essentially reflect the two sections *Carthamus* and *Atractylis* of Bowles et al. [8] although in our analysis *C. glaucus* groups with the sect. *Carthamus* species while in their analysis as well as in Vilatersana et al. [25] and Tarikahya-Hacioglu et al. [35] *C. glaucus* is a member of the sect. *Atractylis* taxa. This inconsistency is however not unique to our analysis. Already in Sasanuma et al. [29], *C. glaucus* is grouping with *C. tinctorius* while Sehgal et al. [12] and Mehrotra et al. [30] retrieved *C. glaucus* accessions in both clades of *Carthamus*. Unfortunately, we relied on gene bank passport data and had no additional seeds available when we obtained the GBS results so that we could not check chromosome numbers for our *C. glaucus* individuals to be sure that they are indeed belonging to the $2n = 20$ -chromosome taxa. As in all before-mentioned studies *C. glaucus* materials from gene banks were used instead of freshly collected plants from the wild, it might be worthwhile to re-determine gene bank-derived *C. glaucus* accessions in future analyses to be sure about the species affiliation of these lineages.

In our second GBS dataset we included a diverse set of 114 *C. tinctorius* accessions together with the sect. *Carthamus* species of our initial analysis but excluded the two obviously introgressed *C. oxyacantha* accessions. While *C. glaucus* appears again paraphyletic in the resulting phylogenetic tree (Fig. 2), the individuals of *C. palaestinus* fall within the safflower accessions occurring in two of their different clades. This result shows that the genomic diversity of *C. tinctorius* cannot be separated from *C. palaestinus* and is a strong indication that the latter is the wild progenitor of safflower, as its genetic diversity is well within the range of *C. tinctorius*. This result is supported by the population assignment analysis where *C. palaestinus* has the highest similarity with *C. tinctorius*, and confirms earlier findings regarding the ancestor of safflower [3,9,12].

Within safflower we found no clear subgroups. *Carthamus tinctorius* in the tree provided in Fig. 2 has four groups of genetically similar accessions, however, only two of the groups are monophyletic and none of them shows a clear geographic pattern. Expectedly, bootstrap values within this species are low, often not even reaching 50%. We interpret this result as an indication of strong trading connections among the early centers of safflower cultivation in Eurasia followed by multiple independent introductions of safflower germplasm from these centers into regions like the Americas and Australia where safflower cultivation started only during the last century. In addition to the mere introduction of foreign germplasm to new areas also gene flow between newly attained and older local cultivars can further blur phylogeographic patterns. This general picture does not change when *C. palaestinus*, *C. oxyacantha* and

C. glaucus individuals with traces of safflower introgression (Fig. 3) are removed from the dataset: In this case the resulting clades and subclades in the phylogenetic analysis again lack geographic structure (see Supplementary Fig. S1 online). This result is in strong contrast to the morphology-defined geographic groups of Kupcov [19], Hanelt [26], Ashri [20] and Knowles and Ashri [1], which all postulated the presence of certain ecotypes typical for specific geographic regions, although they did not arrive at a conclusion about the number of such safflower ecotypes. This discrepancy can be explained either by the quantitative nature of the used morphological traits that are (i) controlled by very few genetic loci with the respective allelic composition not reflected in the GBS data we obtained, and/or (ii) these traits might have an environmental and ecological component that influences characters like flower color or plant height but genetic inheritance might be low. Still, in contrast to our results here Chapman et al. [9], using 24 nuclear microsatellite markers, found five weakly supported groups with a possible geographic correlation. In contrast to our dataset they obtained these groups by restricting their analysis to Ashri's [20] set of assumed autochthonous accessions, i.e. sets of materials that were not influenced by introductions during the last century. For the safflower accessions currently stored in gene banks we conclude that neither geographic origin nor certain morphological traits are usable indicators of genetic similarity or diversity. For breeding approaches, it is therefore necessary to either use molecular analyses to identify rather diverse genotypes or use gene bank passport data to identify accessions with the preferred or extreme agronomic traits.

Conclusions

Using GBS that combines next-generation sequencing with a genomic complexity reduction approach, we were easily able to discern all analyzed safflower accessions and could resolve their phylogenetic relationships. However, within *C. tinctorius* we could not recover any geographically defined groups/ecotypes that were postulated earlier based on morphological character combinations thought to prevail in specific regions. This we explain with seed exchange among the areas where safflower is commercially grown, which might have blurred an eventually formerly existing geographic pattern of locally evolved genotypes.

In contrast to the intra-species structure in *C. tinctorius* our GBS analysis recovered species relationships in *C.* sections *Carthamus* and *Atractylis* with high support values. We identified *C. palaestinus* as closest relative and likely progenitor of safflower. *Carthamus glaucus* resulted as a paraphyletic species that harbors *C. tinctorius*/*C. palaestinus*, but its phylogenetic position in our trees casts some doubts about the species affiliation of the used material. The other analyzed species appear to be monophyletic.

Materials And Methods

Taxon sampling

One hundred sixty-seven individuals (see Supplementary Table S1 online) including 114 cultivated *C. tinctorius* and 53 individuals of wild species of *Carthamus*, i.e. *C. lanatus* (18), *C. oxyacantha* (8), *C.*

glaucus (6), *C. palaestinus* (7), *C. tenuis* (6), and *C. boissieri* (8), were included in GBS-based phylogenetic analyses. A set of eight diverse individuals of *C. tinctorius* was used together with individuals of *C. palaestinus*, *C. oxyacantha* and *C. glaucus* for population structure analyses.

The analyzed materials were obtained as seeds from the gene bank collections of RTIPP-SBUK (Research and Technology Institute of Plant Production, Shahid-Bahonar University, Kerman, Iran), IPK (Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany) and NPGS (National Plant Germplasm System, U.S. Department of Agriculture, USA) and grown at least to a four-leaf stage before harvesting leaves for DNA extraction. Detailed information such as passport data and observation records for each accession can be found in the online GRIN (Germplasm Resources Information Network) database of NPGS (<http://www.ars-grin.gov/npgs/>) and Genebank Information System of the IPK Gatersleben (<https://gbis.ipk-gatersleben.de/>). Collection and use of plant materials was conducted in accordance with institutional, national and international regulations. Seed exchange followed the Nagoya rules of access and benefit sharing through standard material transfer agreements.

DNA extraction and genotyping-by-sequencing (GBS)

Genomic DNA was extracted from young leaves with a DNeasy Plant kit (QIAGEN). To obtain genome-wide SNPs, we used a two-enzyme GBS protocol that uses the restriction enzymes *Pst*I (NEB Inc.) and *Msp*I (NEB Inc.) to digest 200 ng of genomic DNA. For the library preparation and individual barcoding, the protocol of Wendler et al. [36] was followed. The individual libraries were sequenced in multiplex on an Illumina HiSeq 2500 (100 bp single-end reads) at the sequencing facilities of Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany.

GBS data preparation and filtering

Quality assessment of all raw sequence samples was performed using fastqc (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>). Sequences were de-multiplexed allowing for one base mismatch in barcodes, and the restriction site and barcode were trimmed from all GBS sequence reads using cutadapt [37]. Sequence data were filtered and clustered using the ipyrad bioinformatics pipeline [38]. We conducted two runs of ipyrad: (i) one with a small set of *C. tinctorius* accession but including seven wild *Carthamus* species, and (ii) using 114 individuals from *C. tinctorius* plus its closest three relatives according to the prior analysis. The minimal number of samples to possess a certain locus was set to 90% of the individuals. The clustering threshold of reads within and between individuals was set to 0.85 and 0.90 in the first and second dataset, respectively. For the other parameters the default settings were used. The assembly of the GBS data was done de novo.

Phylogenomic analyses

After removing samples with a too low number of reads, we used MrBayes 3.2.6 [39] and Paup* 4a169 [40] to infer phylogenetic relationships within our set of *Carthamus* species and accessions. NJ, based on GTR distances, and MP trees were calculated in Paup* to analyze the aligned sequence data matrices. In

MP gaps were treated as missing data and we used a heuristic search with TBR branch swapping with initially 250 random addition sequences (RAS) restricting the number of collected trees to 50 per replicate. The resulting shortest trees were then used as starting trees in a second heuristic search without setting a limit on tree numbers. Bootstrap support values were obtained by 500 bootstrap re-samples using the same settings as before except that we excluded RAS. Paup* was also used to infer the model of sequence evolution.

For BI two times four chains were run for 5 million generations for the dataset including the set of different *Carthamus* species, specifying the model of sequence evolution as GTR+ Γ . We sampled a tree every 1000 generations and summarized the trees in MrBayes. Converging log-likelihoods, potential scale reduction factors for each parameter and inspection of tabulated model parameters in MrBayes suggested that stationary had been reached. The first 25% of trees were discarded as burn-in. For the large dataset including all *C. tinctorius* accessions, phylogenetic inference calculations in MrBayes were run for more than three weeks without converging chains when we cancelled this analysis.

Population structure analysis

The dataset of SNPs generated from ipyrad was used as input for the analysis of population structure for *C. oxyacantha*, *C. palaestinus*, *C. glaucus* and eight *C. tinctorius* cultivars in the R package LEA [41]. The number of subpopulations (K) was assumed to range from one to seven. An admixture analysis [42] using sparse nonnegative matrix factorization (snmf) was used for the estimation of population genetic structure. In addition, a Principal Components Analysis (PCA) was conducted in ipyrad on the same dataset.

Analysis of genome size

To obtain ploidy levels we used flow cytometry to measure genome sizes of *Carthamus* species [43]. For this, we collected a leaf from each of three individuals from eight *C. tinctorius* and four *C. oxyacantha* accessions and three leaves from one *C. palaestinus* individual. Afterwards the leaves were transported to the lab and genome sizes were measured on a CyFlow flow cytometer (Sysmex Partec GmbH) using propidium iodide (PI) as staining reagent (Sigma-Aldrich) according to Jakob et al. [44]. The tomato cultivar 'Stupicke' was used as the size standard (2C DNA content = 1.96 pg [45]).

Declarations

Acknowledgements

The authors thank C. Koch, B. Kränzlin and P. Oswald for technical assistance in the greenhouses or lab, S. König and A. Himmelbach for GBS library preparation and NG sequencing, and the gene banks of IPK, RTIPP-SBUK and USDA for *Carthamus* germplasm. Financial support by IPK is acknowledged.

Author contributions

Designed study: S.S.N., Z.N., F.R.B. Coordinated study: Z.N., S.S.N., F.R.B. Provided data or materials: G.M.N. Performed experiments: S.S.N., Z.N. Analyzed data: Z.N., S.S.N., F.R.B., R.H. The initial manuscript was written by F.R.B., S.S.N., Z.N. All authors contributed to and approved the final version of the manuscript.

Data availability statement

All sequence data are available through the NCBI nucleotide database under accession numbers XXXXXX–YYYYYY (We are in the process of sequence submission; numbers will be added before final manuscript submission)

Additional information

Competing interest

The authors declare no competing interests.

References

1. Knowles, P. F. & Ashri, A. Safflower: *Carthamus tinctorius* (Compositae) in *Evolution of crop plants*, 2nd. edition (eds. Smartt, J. & Simmonds, N.) 47-50 (Longman, 1995).
2. van Zeist, W. & Walterbolk-van Rooijen, W. Two interesting floral finds from third millennium B.C. Tell Hamman et-Turkman, northern Syria. *Veg. Hist. Archaeobot.* **1**, 157-161 (1992).
3. Chapman, M. A. & Burke, J. M. DNA sequence diversity and the origin of cultivated safflower (*Carthamus tinctorius* L.; Asteraceae). *BMC Plant Biol.* **7**, 1-9 (2007).
4. Bowers, J. E., Pearl, S. A. & Burke, J. M. Genetic mapping of millions of SNPs in safflower (*Carthamus tinctorius* L.) via whole-genome resequencing. *G3* **6**, 2203-2211 (2016).
5. Ali, F. *et al.* Mobile genomic element diversity in world collection of safflower (*Carthamus tinctorius* L.) panel using iPBS-retrotransposon markers. *PLoS One* **14**, e0211985 (2019).
6. Ashri, A. & Knowles, P. F. Cytogenetics of safflower (*Carthamus* L.) species and their hybrids. *Agronomy J.* **52**, 11-17 (1960).
7. Vilatersana, R., Brysting, A. K. & Brochmann, C. Molecular evidence for hybrid origins of the invasive polyploids *Carthamus creticus* and *C. turkestanicus* (Cardueae, Asteraceae). *Mol. Phylogenet. Evol.* **44**, 610-621 (2007).
8. Bowles, V. G., Mayerhofer, R., Davis, C., Good, A. G. & Hall, J. C. A phylogenetic investigation of *Carthamus* combining sequence and microsatellite data. *Plant Syst. Evol.* **287**, 85-97 (2010).
9. Chapman, M. A., Hvala, J., Strever, J. & Burke, J. M. Population genetic analysis of safflower (*Carthamus tinctorius*; Asteraceae) reveals a Near Eastern origin and five centers of diversity. *Am. J. Bot.* **97**, 831-840 (2010).

10. Castañeda-Álvarez, N. P. *et al.* Global conservation priorities for crop wild relatives. *Nat. Plants***2**, 1-6 (2016).
11. Pearl, S. A., Bowers, J. E., Reyes-Chin-Wo, S., Michelmore, R. W. & Burke, J. M. Genetic analysis of safflower domestication. *BMC Plant Biol.* **14**, 1-5 (2014).
12. Sehgal, D., Raina, S. N., Devarumath, R. M., Sasanuma, T. & Sasakuma T. Nuclear DNA assay in solving issues related to ancestry of the domesticated diploid safflower (*Carthamus tinctorius* L.) and the polyploid (*Carthamus*) taxa, and phylogenetic and genomic relationships in the genus *Carthamus* L. (Asteraceae). *Mol. Phylogenet. Evol.* **53**, 631-644 (2009).
13. Mokhtari, N., Rahimmalek, M., Talebi, M. & Khorrami, M. Assessment of genetic diversity among and within *Carthamus* species using sequence-related amplified polymorphism (SRAP) markers. *Plant Syst. Evol.* **299**, 1285-1294 (2013).
14. Agrawal, R., Tsujimoto, H., Tandon, R., Rao, S. R. & Raina, S. N. Species-genomic relationships among the tribasic diploid and polyploid *Carthamus* taxa based on physical mapping of active and inactive 18S–5.8S–26S and 5S ribosomal RNA gene families, and the two tandemly repeated DNA sequences. *Gene***521**,136-144 (2013).
15. Ambreen, H. *et al.* Development of genomic microsatellite markers in *Carthamus tinctorius* L. (safflower) using next generation sequencing and assessment of their cross-species transferability and utility for diversity analysis. *PLoS One***10**, 1-22 (2015).
16. Espanani, S., Majidi, M. M., Saeidi, G., Alaei, H. & Rezaei, V. Wide hybridization and introgression breeding in safflower: Effectiveness of different selection methods. *Plant Breed.* **138**, 846-861 (2019).
17. Singh, K. N. *et al.* Identification and validation of in silico mined polymorphic EST-SSR for genetic diversity and cross-species transferability studies in safflower. *J. Plant Biochem. Biotechnol.* **7**, 1-10 (2021).
18. Elshire, R. J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One***6**, e19379 (2011).
19. Kupcov, A. I. The geographical variation of *Carthamus tinctorius* L. *Bull. Appl. Bot. Genet. Pl. Breed.* **9**, 99-181 (1932). [in Russian]
20. Ashri, A. Evaluation of germplasm collection of safflower, *Carthamus tinctorius* L. V. Distribution and regional divergence for morphological characters. *Euphytica***24**, 651-659 (1975).
21. Johnson, R. C., Kisha, T. J. & Evans, M. A. Characterizing safflower germplasm with AFLP molecular markers. *Crop Sci.* **47**, 1728-1736 (2007).
22. Amini, F., Saeidi, G. & Arzani, A. Study of genetic diversity in safflower genotypes using agro-morphological traits and RAPD markers. *Euphytica***163**, 21-30 (2008).
23. Sehgal, D. & Raina, S. N. Genotyping safflower (*Carthamus tinctorius* L.) cultivars by DNA fingerprints. *Euphytica***146**, 67-76 (2005).
24. Lee, G. A. *et al.* Genetic assessment of safflower (*Carthamus tinctorius* L.) collection with microsatellite markers acquired via pyrosequencing method. *Mol. Ecol. Resour.* **14**, 69-78 (2014).

25. Vilatersana, R., Susanna, A., Garcia-Jacas, N. & Garnatje, T. Generic delimitation and phylogeny of the *Carduncellus*-*Carthamus* complex (Asteraceae) based on ITS sequences. *Plant Syst. Evol.* **221**, 89-105 (2000).
26. Hanelt, P. Zur Kenntnis von *Carthamus tinctorius* L. *Kulturpflanze* **9**, 114-145 (1961).
27. Vilatersana, R., Susanna, A., Garcia-Jacas, N. & Garnatje, T. Karyology, generic delineation and dysploidy in the genera *Carduncellus*, *Carthamus* and *Phonus* (Asteraceae). *Bot. J. Linn. Soc.* **134**, 425-438 (2000).
28. Sehgal, D., Rajpal, V. R. & Raina S. N. Chloroplast DNA diversity reveals the contribution of the two wild species in the origin and evolution of diploid safflower (*Carthamus tinctorius* L.). *Genome* **51**, 638-643 (2008).
29. Sasanuma, T., Sehgal, D., Sasakuma, T. & Raina, S.N. Phylogenetic analysis of *Carthamus* species based on the nucleotide sequence of the nuclear SACPD gene and chloroplast *trnL-trnF* IGS region. *Genome* **51**, 721-727 (2008).
30. Mehrotra, S., Goel, S., Sharma, S., Raina, S. N. & Rajpal, V. R. Sequence analysis of *KpnI* repeat sequences to revisit the phylogeny of the genus *Carthamus* L. *Appl. Biochem. Biotechnol.* **169**, 1109-1125 (2013).
31. Anderson, B. M., Thiele, K. R., Krauss, S. L. & Barrett, M. D. Genotyping-by-sequencing in a species complex of Australian Hummock Grasses (*Triodia*): Methodological insights and phylogenetic resolution. *PLoS One* **12**, e0171053 (2017).
32. Milner, S. G. *et al.* Genebank genomics highlights the diversity of a global barley collection. *Nat. Genet.* **51**, 319-326 (2019).
33. Nemati, Z., Harpke, D., Gemicioglu, A., Kerndorff, H. & Blattner, F. R. Saffron (*Crocus sativus*) is an autotriploid that evolved in Attica (Greece) from wild *Crocus cartwrightianus*. *Mol. Phylogenet. Evol.* **136**, 14-20 (2019).
34. Bernhardt, N. *et al.* Genome-wide sequence information reveals recurrent hybridization among diploid wheat wild relatives. *Plant J.* **102**, 493-506 (2020).
35. Tarıkahya-Hacıoğlu, B., Karacaoğlu, Ç. & Özüdoğru, B. The speciation history and systematics of *Carthamus* (Asteraceae) with special emphasis on Turkish species by integrating phylogenetic and ecological niche modelling data. *Plant Syst. Evol.* **300**, 1349-1359 (2014).
36. Wendler, N. *et al.* Unlocking the secondary gene-pool of barley with next-generation sequencing. *Plant Biotech. J.* **12**, 1122-1131 (2014).
37. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10-12 (2011).
38. Eaton, D. A. PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* **30**, 1844-1849 (2014).
39. Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539-542 (2012).

40. Swofford, D. L. *PAUP*.Phylogenetic analysis using parsimony (* and other methods)*, v4.0a 169 (Sinauer Assoc., 2002).
41. Frichot, E. & François, O. LEA: An R package for landscape and ecological association studies. *Methods Ecol. Evol.* **6**, 925-929 (2015).
42. Patterson, N., Price, A. L. & Reich D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
43. Garnatje, T., Garcia, S., Vilatersana, R. & Vallès, J. Genome size variation in the genus *Carthamus* (Asteraceae, Cardueae): Systematic implications and additive changes during allopolyploidization. *Ann. Bot.* **97**, 461-467 (2006).
44. Jakob, S. S., Meister, A. & Blattner, F. R. The considerable genome size variation of *Hordeum* species (Poaceae) is linked to phylogeny, life form, ecology, and speciation rates. *Mol. Biol. Evol.* **21**, 860-869 (2004).
45. Doležal, J. & Greilhuber, J. Nuclear genome size: Are we getting closer? *Cytometry* **77A**, 635-642 (2010).

Table

Table 1. Genome sizes of *C. tinctorius*, *C. oxyacantha*, and *C. palaestinus* individuals measured by flow cytometry against tomato cv. 'Stupické' (2C = 1.96 pg) as size standard.

Accession	Species	Genome size pg/2C (± SD)
C_tin_TN79548_Hungary	<i>C. tinctorius</i>	2.50 (±0.03)
C_tin_CW74_USA	<i>C. tinctorius</i>	2.59 (±0.05)
C_tin_CART156_Pakistan	<i>C. tinctorius</i>	2.49 (±0.05)
C_tin_CART86_Tunisia	<i>C. tinctorius</i>	2.46 (±0.02)
C_tin_CART118_India	<i>C. tinctorius</i>	2.44 (±0.06)
C_tin_CART227_Germany	<i>C. tinctorius</i>	2.50 (±0.02)
C_tin_K101_Iran	<i>C. tinctorius</i>	2.54 (±0.03)
C_tin_Goldasht_Iran	<i>C. tinctorius</i>	2.68 (±0.09)
C_oxy_IR_Aligoodarz	<i>C. oxyacantha</i>	2.54 (±0.02)
C_oxy_IR_Azerbaijan	<i>C. oxyacantha</i>	2.53 (±0.04)
C_oxy_IR_Hamedan	<i>C. oxyacantha</i>	2.40 (±0.07)
C_oxy_IR_Isfahan	<i>C. oxyacantha</i>	2.42 (±0.05)
C_pal_PI235663_1	<i>C. palaestinus</i>	2.60 (±0.01)

Figures

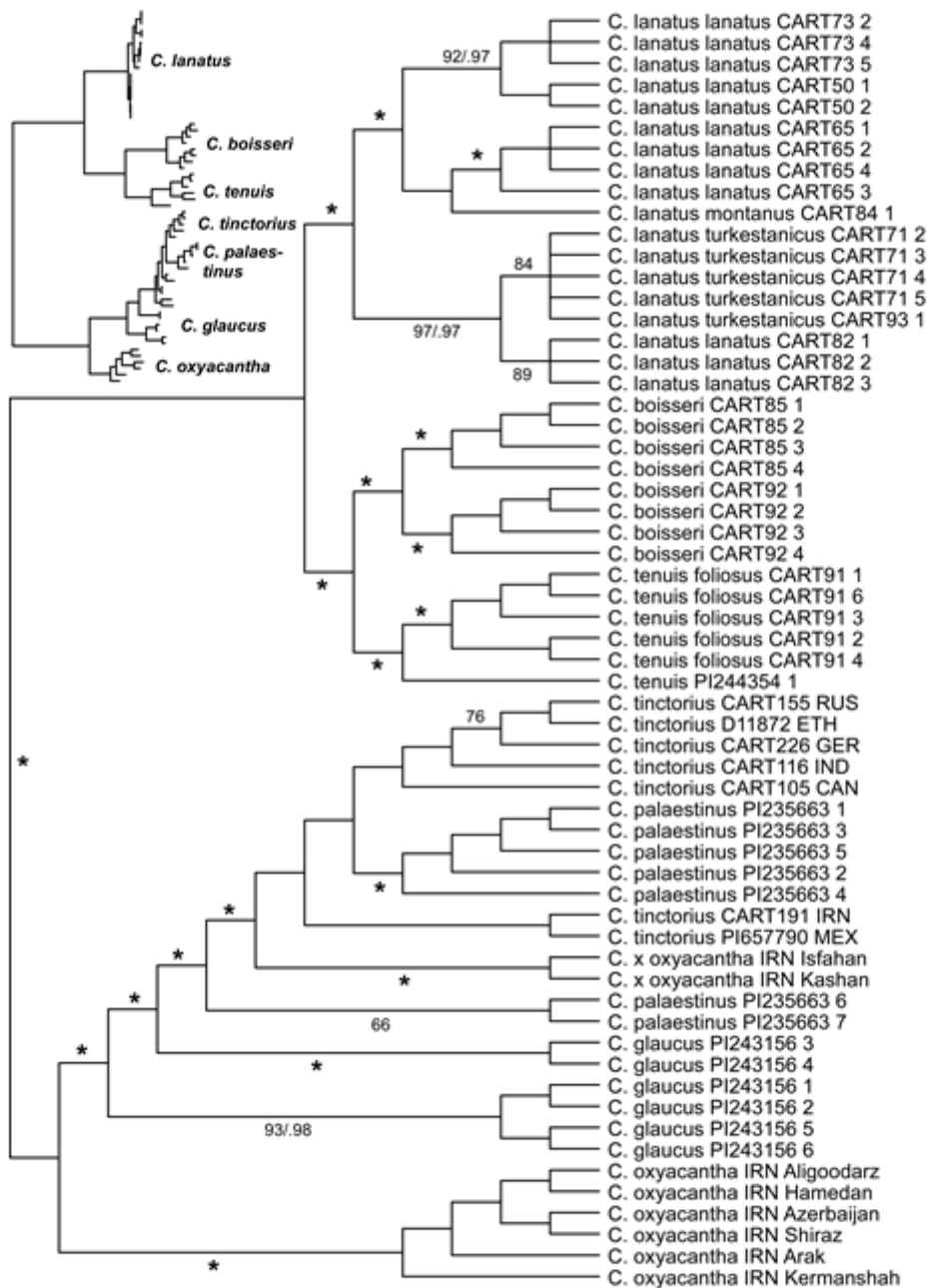


Figure 1

Unrooted strict consensus tree of 2484 MP trees. Asterisks indicate branches with support values of $\geq 98\%$ in MP bootstrap analysis and posterior probabilities ≥ 0.99 in Bayesian phylogenetic inference for the backbone clades of the tree. The inserted tree on the left is one of the most parsimonious trees and indicates branch lengths in MP analysis.

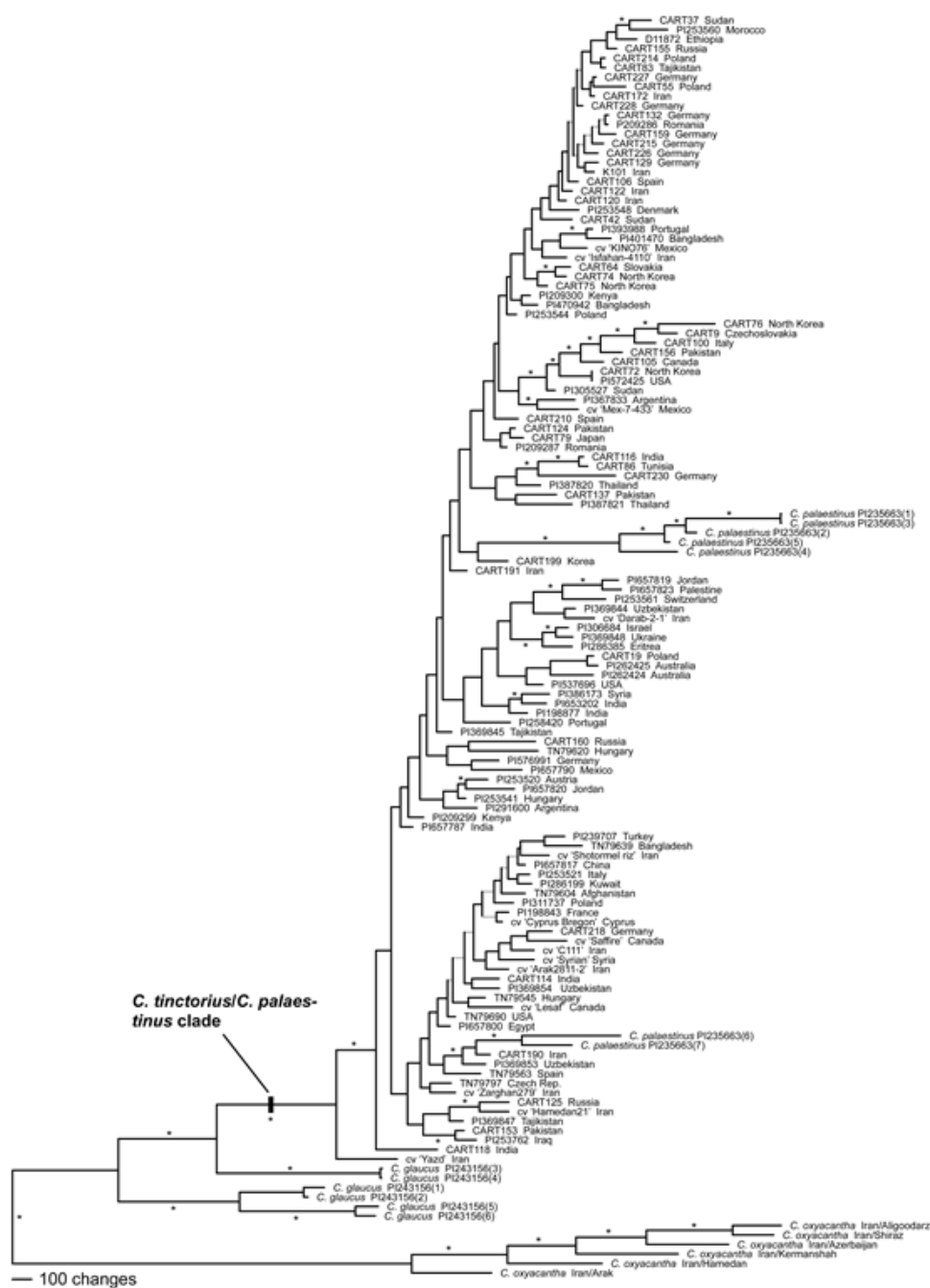


Figure 2

One out of four equally parsimonious trees of a GBS-derived MP analysis of a large and geographically diverse set of 114 safflower accessions. The tree was rooted with *C. oxyacantha*. Bootstrap values $\geq 75\%$ are indicated by asterisks at the branches. Gray branches depict clades that were not recovered in all MP trees and collapse in the strict consensus tree.

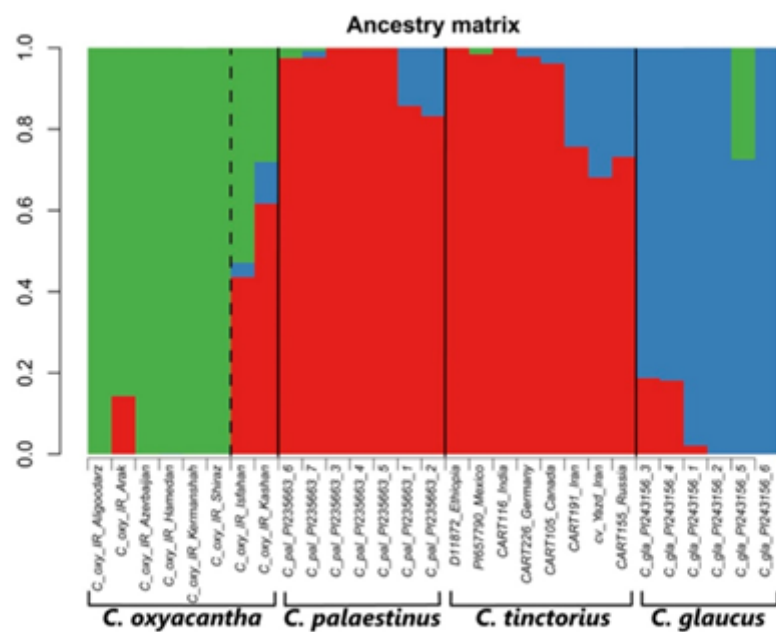


Figure 3

Population structure analysis with $K = 3$ based on 3720 unlinked SNPs of, from left to right, *C. oxyacantha*, *C. palaestinus*, *C. tinctorius* and *C. glaucus*.

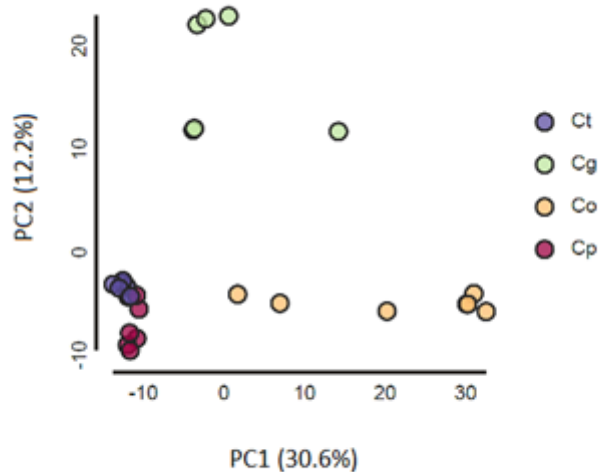


Figure 4

Principal Component Analysis (PCA) of *C. oxyacantha* (Co), *C. palaestinus* (Cp), *C. glaucus* (Cg) and *C. tinctorius* (Ct), giving distributions on the first and second PCs.

Supplementary Files

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

- [SupplementaryInformation.pdf](#)