Genome Biology

RESEARCH

Open Access

Pan-genome analysis of *Caragana* reveals genomic mechanisms of drought adaptation of leguminous shrubs to arid environments

Jiaxing Song[1†], Jie An[1†], Houyin Xia[1], Wanjiang Li[1], Jingdian Hu[1], Qian Li[2], Chenjian Li[2], Xiaoxia Shen[1], Hong Cheng[1], Yi Zhou[3], Tianming Hu[1] and Peizhi Yang[1*]

[†]Jiaxing Song and Jie An contributed equally to this work.

*Correspondence: yangpeizhi@126.com

[1] College of Grassland Agriculture, Northwest A&F University, Yangling, Shaanxi 712100, China
[2] Key Laboratory of Ministry of Education of Grassland Resources and Ecology of Western Arid Region, Key Laboratory of Grassland Resources and Ecology of Xinjiang, College of Grassland Science, Xinjiang Agricultural University, Urumqi 830052, China
[3] School of Agriculture, Food and Wine, The University of Adelaide, Urrbrae, SA 5064, Australia

Abstract

Background: Perennial leguminous shrubs of the genus *Caragana* play key roles in arid and semi-arid ecosystems, contributing to ecological restoration, nitrogen fixation, and bioenergy production. However, the genomic basis underlying their adaptation to arid environments remains insufficiently understood.

Results: Here, we generate chromosome-scale genome assemblies for 14 representative *Caragana* species, including a near telomere-to-telomere assembly of *Caragana korshinskii* (NMCN13). Using these genomes, we construct a graph-based pan-genome and identify extensive genomic variation, including 801,351 structural variants and 104,404 presence-absence variations. Comparative analyses identify clade-specific gene family expansions enriched in stress-related pathways, particularly secondary metabolism. We further identify two candidate drought-responsive genes, *CkoERF1* and *CkoWAKL*, whose functions are supported by heterologous overexpression in *Arabidopsis thaliana* and alfalfa, demonstrating enhanced drought-response phenotypes. Transposable element analyses indicate that variation in LTR retrotransposon abundance contributes to genome size diversity across species. Notably, we find that some key symbiosis-related genes, for example, *NIN*, *DMI1*, *DMI2*, and *NSP1*, are conserved across all species, suggesting a stable genetic basis for nitrogen fixation.

Conclusions: This study provides high-quality genomic resources and a comprehensive pan-genomic framework for *Caragana*. Our findings offer new insights into genome evolution, environmental adaptation, and drought tolerance mechanisms, and establish a foundation for future functional genomics and breeding efforts in leguminous shrubs.

Keywords: Pan-genome, *Caragana*, Arid adaptation, Structural variants, Drought stress

Background

The Fabaceae family, which includes agriculturally important legumes, is widely distributed across diverse ecosystems around the world [1]. In addition to being important to agricultural settings, the Fabaceae family plays critical roles in ecosystem restoration.

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

The nitrogen fixing capacity of legumes improves the quality of nutrient-poor soils [2]. Their combined ecological benefits and economic value make legumes very important plants [3]. However, the rising incidences of drought, exacerbated by climate change, threaten the global agricultural production of such vital legumes. The International Union for Conservation of Nature (IUCN) has sounded the alarm on the urgent need for a coordinated and comprehensive approach to address drought, desertification, and land degradation to protect the planet's ecosystem stability and people's livelihoods [4, 5]. The genus *Caragana* comprises perennial leguminous shrubs that are widely distributed across deserts, semi-deserts, grasslands, and shrublands in the arid regions of China. These species play important roles in soil and water conservation, windbreak establishment, sand fixation, bioenergy production, and nitrogen fixation, and they also exhibit remarkable tolerance to drought, cold, and heat stress [6–8]. Like many temperate legumes, *Caragana* species can establish symbiotic associations with soil bacteria, particularly rhizobia within the genera *Mesorhizobium* and *Rhizobium*, through the formation of determinate nodules [9, 10]. It is noteworthy that drought adaptation and species diversification in the genus *Caragana* have become research hotspots [11, 12]. However, pangenomic resources for the genus *Caragana* remain scarce, which hinders comprehensive comparative analyses and investigations into its stress tolerance mechanisms.

The recent and rapid development of high-throughput sequencing technologies has accelerated comparative genomic research in Fabaceae [12, 13]. Comparative genomics and analysis of structure, function, and evolutionary relationship of the genomes of different Fabaceae species have revealed the molecular mechanisms of symbiotic nitrogen fixation, species diversity, and adaptive evolution [14, 15]. While reference genomes serve as important resources for understanding gene structure and function within a species or lineage, reliance on a single reference genome may introduce reference bias and lead to information loss, particularly for sequences that are absent from, highly diverged from, or structurally rearranged relative to the chosen reference haplotype [16–18]. Pan-genomics effectively mitigates the biases and limitations of a single reference genome [19]. By comprehensively integrating the genetic diversity for a species as fully as possible, pan-genomes provide deeper insights into population genetics, evolutionary dynamics, and genomic complexity beyond traditional single-reference approaches [20]. Pan-genomic approaches have been widely applied in agriculturally important crops such as rice [21], maize [22], soybean [23], barley [24], poplar [25], and tea [26]. For instance, the analysis of pearl millet pan-genomes revealed the presence of structural variants (SVs) that contribute to heat tolerance by influencing gene regulation [27]. Further insights may be gained by better understanding the various important roles of SVs in plant growth and in responding to biotic and abiotic stresses.

In this study, we generated chromosome-scale reference genomes for 14 representative *Caragana* species and constructed a graph-based pan-genome to investigate interspecific genomic variation across the genus. Pan-genome analysis revealed that the highly conserved genes *DMI1* (Does Not Make Infections 1) and *NIN* (Nodule Inception), which are essential for driving symbiotic associations in Fabaceae, are also well-preserved in *Caragana*. Additionally, we identified extensive structural and presence–absence variation, examined the conservation of symbiosis-related genes, and prioritized two drought-related candidate genes, *CkoERF1* and *CkoWAKL*, for functional

analysis. Transcriptomic analysis and heterologous functional assays further supported drought-response roles for the prioritized candidate genes under experimental conditions. Our study provides a comparative genomic analysis of *Caragana* species and offers a useful framework for investigating their distribution patterns and genomic correlates of adaptation to extreme environments.

Results

De novo assembly and annotation of multiple ecologically diverse *Caragana* species

Caragana korshinskii (NMCN13) is adapted to extremely arid environments (Additional file 1: Fig. S1a). Its ability to prevent soil erosion and reverse desertification has made it the second most planted species as a part of China's Three-North Shelter Forest Programme to promote afforestation [28] (Additional file 1: Fig. S1b and S1c). In order to understand the genomic basis of NMCN13 adaptation to extreme drought, we assembled the near telomere-to-telomere (near-T2T) reference genome for NMCN13 by integrating 71.496 Gb of PacBio HiFi data and 62.918 Gb of ONT data (Additional file 2: Table S1), together with 107.037 Gb of Hi-C data and 282.376 Gb of Illumina paired-end sequences (Additional file 2: Table S2). The analysis of telomeric repeats (AAACCCT) in our assembly confirmed that all 16 telomeres were successfully detected, indicating that the genome assembly reached the near-T2T level of completeness (Additional file 2: Table S3), while the predicted centromere coordinates for the eight NMCN13 chromosomes are provided in Additional file 2: Table S4. To elucidate the phylogenetic relationships among the assembled *Caragana* species, we identified conserved single-copy orthologous genes across the 14 newly assembled genomes and representative related taxa and reconstructed a genome-scale species phylogeny. This ortholog-based phylogeny was used as the comparative framework for subsequent analyses of gene-family evolution, pan-genome composition, and structural variation.

To further capture genomic diversity within the focal group analyzed in this study, we selected 13 additional *Caragana* species from different regions of China on the basis of their phylogenetic positions and ecological representativeness, together with *C. korshinskii* (NMCN13), for de novo genome sequencing and comparative analysis (Fig. 1A; Additional file 1: Fig. S2; Additional file 2: Table S5). These 13 *Caragana* species were sequenced using PacBio reads to generate approximately 41.053–70.851 Gb (Additional file 2: Table S6). Across all 14 accessions, Hi-C sequencing yielded 100.05–159.171 Gb per accession (Additional file 2: Table S7), whereas the Illumina paired-end short-read datasets used for genome survey contained 28.71–56.27 Gb per accession (Additional file 2: Table S8). Just prior to assembly, we generated a frequency distribution of $K = 21$ k-mers using short-read sequences and estimated genome sizes, and the results showed that the genome size was 0.93–1.36 Gb and the heterozygosity rate was 0.83–1.81% (Additional file 1: Fig. S3; Additional file 2: Table S9).

We performed the preliminary assembly using PacBio reads, followed by three rounds of correction and polishing utilizing short-read data to remove redundancy and improve accuracy. The final genome sizes assembled ranged from 0.931 to 1.384 Gb, with contig N50 values ranging from 48.0 to 163.72 Mb (Fig. 1B; Additional file 2: Table S9). Repetitive sequences comprised 49.6–65.0% of the assemblies (Fig. 1B). Subsequent high-throughput chromatin conformation capture (Hi-C) generated eight chromosome-scale

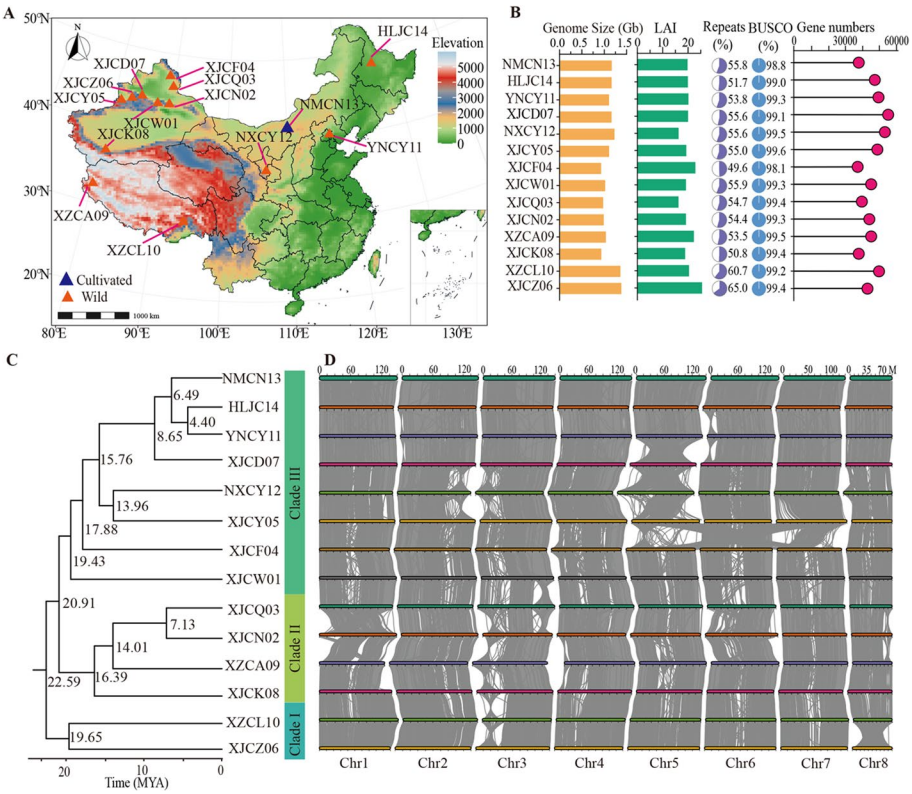


Fig. 1 The 14 *Caragana* sources and genome assembly. **A** Habitat and phenotype of 14 *Caragana* species, including *C. acanthophylla* (XJCW01), *C. pumila* (XJCN02), *C. altaica* (XJCQ03), *C. tragacanthoides* (XJCF04), *C. leucospina* (XJCY05), *C. jubata* (XJCZ06), *C. soongorica* (XJCD07), *C. polourensis* (XJCK08), *C. versicolor* (XZCA09), *C. bicolor* (XZCL10), *C. pekinensis* (YNCY11), *C. roborovskyi* (NXCY12), *C. korshinskii* (NMCN13), and *C. arborescens* (HLJC14). The geographical map was adapted from the one provided by the Data V.GeoAtlas (https://datav.aliyun.com/portal/school/atlas/area_selector). **B** Genome assembly and annotation of 14 *Caragana* species, including genomic length (Gb), Long terminal repeat assembly index (LAI), repeat percentage, benchmarking universal single-copy orthologs (BUSCO) and gene number. **C** Genome-scale phylogenetic relationships among the 14 assembled *Caragana* species inferred from orthologous gene families in a broader comparative genomic framework. Different colors indicate the three principal clades recovered in this analysis. **D** Synteny plot across the 14 *Caragana* species

pseudomolecules at the chromosome level. The final reference scaffold assemblies spanned from 0.834 to 1.315 Gb, representing approximately 89.23% to 98.36% of the assembled genome size (Additional file 1: Fig. S4; Additional file 2: Table S10). Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis showed that 98.14–99.57% of the BUSCO sequences were completely present in the *Caragana* genomes (Fig. 1B; Additional file 2: Table S10), indicating the high completeness of our assemblies. To evaluate the accuracy of the assembled genomes, we aligned paired-end short reads against the assembled genomes. The consensus quality value (QV) scores ranged from 37.34 to 49.31, confirming the high base-level accuracy of the assemblies (Additional file 2: Table S11). Additionally, the calculated Long Terminal Repeat Assembly Index (LAI) values ranged from 16.24 to 25.07 across the 14 genomes (Fig. 1B; Additional file 2: Table S11).

The telomeric repeat sequence AAACCCT was detected on the telomeres of all chromosomes in the NMCN13 genome assembly. Notably, only one gap was identified in

Chromosome 8 (Chr8), indicating that the genome assembly had achieved near-T2T resolution. Furthermore, the distribution of telomeric repeat signals, assembly gaps, predicted centromeric regions, and predicted 45S precursor rRNA loci across all 14 *Caragana* genome assemblies is summarized in Additional file 1: Fig. S5. In order to achieve comprehensive and high-quality gene annotation, we used a federated evidence-based gene prediction strategy and combined orthologous proteins, ab initio predictions, and transcriptome data. RNA-seq quality and mapping statistics supporting gene annotation for the 14 accessions are summarized in Additional file 2: Table S12. Upon integrating the predictions from these three methods, we generated 37,661 to 56,234 protein-coding genes per assembly (Fig. 1B; Additional file 2: Table S13). BUSCO evaluation of these protein-coding genes in each assembly resulted in scores ranging from 93.3% to 99.7% and an average score of 95.24% (Additional file 2: Table S13). These metrics demonstrated that the 14 assembled genomes in this study exhibited high contiguity and completeness and were reliable to use for further downstream analyses.

Phylogenetic analysis and clade-restricted gene-family expansions in *Caragana*

To elucidate the evolutionary history of the genus *Caragana*, we conducted comparative genomic analyses of 31 species, including 16 leguminous plants, 14 *Caragana* species and the model plant *Arabidopsis thaliana*. We performed cluster analysis on these 31 species and identified 6,662 species-specific orthogroups (unique to individual species) using OrthoFinder (v2.5.5), which were utilized to investigate species-specific evolutionary patterns (Additional file 2: Table S14). Upon constructing a phylogenetic tree of the 31 species, we found that the genus *Caragana* is closely related to *Oxytropis ochrocephala*, with a divergence time of approximately 39.83 million years ago (Additional file 1: Fig. S6). To further elucidate the phylogenetic relationships among the assembled *Caragana* species, we identified orthologous gene families across the 14 newly assembled genomes and representative related taxa, and reconstructed a genome-scale species phylogeny (Fig. 1C). The phylogenetic analysis of the 14 *Caragana* species revealed three distinct clades, with two divergence events occurring around 22.59 and 20.91 million years ago, broadly coinciding with the rapid uplift of the Himalayas around 20 million years ago [29] (Fig. 1C). The resulting phylogeny recovered three distinct clades within the sampled *Caragana* species, providing an evolutionary framework for subsequent comparative genomic analyses (Fig. 1C). These results provide a useful framework for evolutionary studies of the genus *Caragana*. A functional domain annotation analysis of expanded genes from the 14 *Caragana* species revealed significant enrichment in Clade I for 40S_S4_c (40S ribosomal protein S4) and PAP_assoc (Cid1 family poly A polymerase), in Clade II for SHS2_Rpb7-N (SHS2 domain of Rpb7p N-terminus), ENTH (Epsin NH2-terminal homology), and DGAT (Diacylglycerol-O-acyltransferase), and in Clade III for Wax2_C (C-terminal domain of Wax2), Glyco_hydro_19 (Glycoside hydrolase 19), and EIF2A (Eukaryotic initiation factor 2A) (Additional file 1: Fig. S7). In addition, we found that AP2/ERF (APETALA2/Ethylene Responsive Factor) was significantly expanded in NMCN13 of Clade III, compared to the other 13 *Caragana* species. The KEGG pathway enrichment analysis showed that the expanded genes were enriched in the monobactam biosynthesis, lysine biosynthesis, indole alkaloid biosynthesis, and

betaine biosynthesis pathways ($P < 0.05$) (Additional file 1: Fig. S8), which may play an important positive role in the environmental adaptability of *Caragana* species.

To investigate whole genome duplication (WGD) events in the genus *Caragana*, we analyzed synonymous substitution rates (K_S) of paralogous gene pairs in syntenic blocks of three *Caragana* species: XJCZ06 (Clade I), XZCA09 (Clade II), and NMCN13 (Clade III). The distributions of K_S for XJCZ06, XZCA09, and NMCN13 showed two distinct peaks: a first peak centered at K_S values of 0.53, 0.55, and 0.57, and a second peak at 1.76, 1.80, and 1.73, respectively (Additional file 1: Fig. S9). The occurrence timing of the first peak around 0.55 was earlier than the divergence of three Papilionoideae species (including *Melilotus officinalis*, *Medicago sativa* and *M. truncatula*), and it has been confirmed that the duplicated genes from this peak were derived from the Papilionoid WGD event [12]. The second peak corresponds to the well-known gamma (γ) whole-genome triploidy (WGT) event common to eudicots [12]. In addition, to further determine whether a unique WGD occurred, we conducted collinearity analysis using the WGDI software, with *M. truncatula* and *Vitis vinifera*, neither of which has undergone a recent lineage-specific whole-genome duplication after the shared ancestral polyploidization events, serving as reference species. Synteny analysis showed that NMCN13 had 1:1 homologous blocks with *M. truncatula*, while no evidence of residual duplication was found in the comparison with *Vitis vinifera* (Additional file 1: Fig. S10). The analysis results indicate that *NMCN13* has not undergone additional lineage-specific WGD following the shared WGD event in the Papilionoideae subfamily. These findings provide a chromosome-scale genomic resource for the genus *Caragana* and establish a comparative framework for investigating interspecific genome evolution, environmental adaptation, and trait diversification.

Pan-genome analyses of ecologically diverse *Caragana* species

To investigate the genomic diversity of the genus *Caragana*, we constructed a pan-genome using 14 *Caragana* species from diverse ecological niches in China (Fig. 1D). With the addition of every accession, the total number of pan gene families increased before plateauing (with the number of newly added pan-gene families increasing by less than 1% as additional genomes were incorporated) (Fig. 2A). In contrast, the number of core gene families decreased as additional genomes were incorporated, a trend that is consistent with the results of the *Arabidopsis* pan-genome study [30]. The pan-genome includes 14,868 core gene families (found in all species), 1,166 private gene families (present in only one species), and 11,772 dispensable gene families (present in 2 to 13 *Caragana* species), which account for 53.5%, 4.2%, and 42.3% of the total pre-clustered gene families, respectively (Fig. 2B and C, Additional file 2: Table S15). KEGG annotation analysis on all gene families revealed a significant enrichment of core genes ($P < 0.05$) in betalain biosynthesis, lipoic acid metabolism, sulfur relay system, photosynthesis, other types of O-glycan biosynthesis, brassinosteroid biosynthesis, thiamine metabolism, glycosylphosphatidylinositol (GPI), and zeatin biosynthesis (Additional file 1: Fig. S11; Additional file 2: Table S16). These results indicate that the core gene set of *Caragana* retains conserved genes associated with fundamental biological processes, including symbiosis-related functions. Enrichment analysis showed that genes from private gene families were significantly overrepresented ($P < 0.05$) in KEGG pathways related to

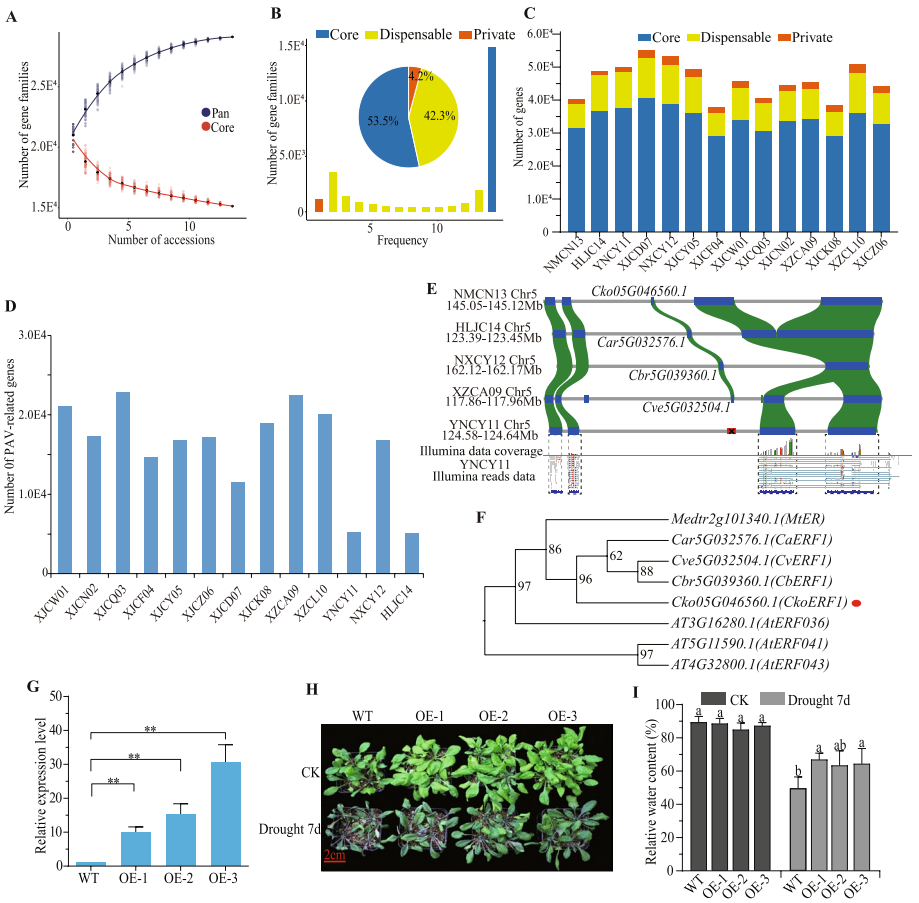


Fig. 2 Pan-genome of 14 *Caragana* species. **A** Modeling of the core genome (red curve) and the pan-genome (blue curve). **B** The number of gene families in the 14 genomes with different frequencies. The pie charts indicate the proportion of gene families for each constituent marker. **C** The number of genes in each composition in individual *Caragana* genomes. **D** Number of PAV-related genes within each genome compared to the NMCN13 reference genome. **E** Schematic representation of structurally distinct variants identifiable between chromosome-level assemblies of the *Caragana* genome. **F** Phylogenetic tree of the ERF gene family members. The tree illustrates the evolutionary relationships among *ERF* orthologs across four *Caragana* accessions and two model plants. **G** Relative expression of *CkoERF1* in transgenic *A. thaliana* overexpression lines. **H** Phenotypes of wild-type (WT) and three independent *CkoERF1* overexpression lines (OE-1, OE-2, and OE-3) under control (CK, well-watered) and drought treatment (Drought-7d, drought conditions for 7 days). **I** Relative water content (RWC) of WT and *CkoERF1* overexpression lines under drought treatment. Data are presented as mean ± SD from independent biological replicates

one-carbon metabolism mediated by folate, C5-branched dibasic acid metabolism, and biotin metabolism. In contrast, genes from dispensable gene families were significantly overrepresented ($P < 0.05$) in pathways related to glucosinolate biosynthesis, other glycan degradation, lysine biosynthesis, riboflavin metabolism, selenocompound metabolism, caffeine metabolism, and indole alkaloid biosynthesis (Additional file 2: Table S16). The enrichment of pathways related to plant secondary metabolism suggests that dispensable genes may be associated with adaptation to local environmental conditions [31].

Core genes play a vital role in maintaining fundamental functions in plants [30, 32]. Because root-nodule symbiosis is a defining feature of many legumes, we examined whether several well-established symbiosis-related genes were retained across the 14 *Caragana* genomes. We focused on four genes with recognized roles in nodulation and

symbiotic signaling in *Medicago truncatula*: *DMI1* [33] (*Does Not Make Infections 1*), *NIN* [34] (*Nodule Inception*), *NSP1* [35] (*Nodulation Signaling Pathway*), and *DMI2* [36] (*Does Not Make Infections 2*), which are well known for their roles in *M. truncatula*. We found that the corresponding four homologous genes in *Caragana*, which we identified through homology comparison analysis, belong to the core gene family (Additional file 1: Fig. S12a-d; Additional file 2: Table S17). *NIN* homologues were identified in all 14 sampled *Caragana* genomes, were classified as core genes, and retained the characteristic PB1 and RWP-RK domains of NIN-like proteins (Additional file 2: Table S17). These observations indicate that homologues of several canonical nodulation and symbiotic-signaling genes are retained as core gene families across the sampled *Caragana* genomes. This pattern supports conservation at the gene-family level and, for *NIN*, at the domain-architecture level, although functional variation in nodulation and symbiotic performance among species remains to be tested experimentally.

Analyses of structural variants (SVs) and presence-absence variants (PAVs) in *Caragana* species

To identify SVs in 14 *Caragana* species, we constructed a pan-genome map using the vg tool by integrating variants from the MUMmer alignment and used NMCN13 as the reference. The resulting pan-genome graph comprised a total of 6.32 Gb with 95,365,780 nodes (the number of fragments of sequences) and 122,449,332 edges (the connections between nodes). We obtained a total of 801,351 SVs, consisting of 104,404 PAVs, 86,071 inversions (INV), 369,324 duplications (DUP), and 241,552 translocations (TRANS) (Additional file 2: Table S18). In plants, inversions and translocations can alter recombination patterns and linkage relationships, thereby affecting the genomic context in which allelic variation is inherited [37]. These SVs, including inversions and translocations, highlight substantial genomic differentiation among *Caragana* clades and provide candidate regions for future investigation of their possible evolutionary significance. Our SV analysis revealed that XJCQ03 and XZCA09 of clade II had the highest number of PAV-related genes (Fig. 2D; Additional file 2: Table S19). This observation likely stems from these two species being the most evolutionarily distinct and divergent in terms of their sequence similarity. Conversely, XJCD07, HLJC14, and YNCY11 are evolutionarily closer to NMCN13 (all members of clade III) and displayed fewer PAV-related genes and accumulated variations compared to the other ten species (Additional file 2: Table S19). We performed enrichment analysis of transcription factor families, protein domains, and stress-related functional categories among PAV-related genes across 13 *Caragana* species (Additional file 1: Fig. S13). The results showed that different *Caragana* species were enriched for distinct regulatory or stress-associated categories. For example, PAV-related genes were enriched for Glutathione S-transferase (GST)-related genes in XJCZ06, zinc finger RING (Zf-RING) domain-containing genes in XZCL10, Homeobox transcription factor genes in XJCK08, and jasmonate-related regulatory genes in XJCN02. Additional enriched categories included AP2 transcription factors, cytochrome P450 enzymes, glyoxalase-related enzymes, and EF-hand domain-containing proteins in other species. These transcription factor families, protein domains, and stress-associated enzymatic categories are involved in plant stress responses and may contribute to environmental adaptation in *Caragana*.

To explore whether dispensable gene-family composition covaried with site-level environmental conditions, we performed redundancy analysis (RDA) using a Hellinger-transformed binary matrix of dispensable gene-family presence/absence across the 14 sampled *Caragana* genomes (Additional file 2: Table S20). The constrained component accounted for 46.72% of the total variation, and the first two constrained axes (RDA1 and RDA2) explained 22.66% and 17.03% of the constrained variation, respectively. In the RDA ordination, average annual precipitation and relative humidity contributed strongly to the constrained axes, suggesting that variation in dispensable gene-family composition covaried with water-availability-related environmental gradients among the sampled accessions (Additional file 1: Fig. S14a and S14b). We further compared transcription factor family annotations among dispensable genes in NMCN13, YNCY11, and HLJC14, which were selected for downstream comparison because their collection sites differed in precipitation and relative humidity. Compared with YNCY11 and HLJC14, dispensable genes in NMCN13 showed a higher representation of AP2/ERF, B3, C2, FAR1, and MYB family members (Additional file 1: Fig. S14c). The AP2/ERF gene family was significantly expanded in NMCN13 but not in YNCY11 (Additional file 1: Fig. S7). The AP2/ERF gene family plays an essential role in regulating biological processes related to promoting plant response to environmental stress and regulating growth [38]. Therefore, we performed genome-wide characterization and analysis of AP2/ERF gene family in 14 *Caragana* species and found 180 AP2 domain-containing genes in NMCN13 (Additional file 2: Table S21). In addition, we identified that a deletion mutation in YNCY11 within the 124.58–124.64 Mb region on Chr5 resulted in the loss of an *ERF* gene. The gene was retained in NMCN13 and HLJC14 (Fig. 2E). Notably, the genes *Cko05G046560.1* (*CkoERF1*) and *Car5G032576.1* (*CarERF1*) of the AP2/ERF gene family were most closely related to *AtERF036* (*AT3G16280.1*) of *A. thaliana* (Fig. 2F; Additional file 1: Fig. S15). Because NMCN13, YNCY11, and HLJC14 differed in collection-site humidity and precipitation, we further evaluated whether the retained *ERF* gene in NMCN13 could influence drought-response phenotypes in a heterologous system (Additional file 1: Fig. S14a). To further investigate the functional significance of *CkoERF1*, we overexpressed this gene in *A. thaliana* and obtained three independent overexpression lines (Fig. 2G). Under seven days of water-withholding treatment, these overexpression lines showed improved drought-response phenotypes and significantly higher relative water content (RWC) than wild-type plants (Fig. 2H and I). These findings support the inference that *CkoERF1* can enhance drought-response phenotypes in a heterologous model system.

Transposable element landscapes in *Caragana*

The amplification, differential retention, and apparent loss of TE-derived sequences can influence genome size and genome evolution in plants [39]. While TEs are recognized as a major driver of genome size variation in plants, their characterization within the genus *Caragana* remains limited. To classify TEs in *Caragana*, we analyzed long terminal repeat (LTR) retrotransposon subfamilies, with a special focus on the Gypsy superfamily, which has been associated with plant genome size and SVs [40]. Our analyses showed that Gypsy retrotransposons were the most abundant LTR retrotransposon superfamily across the 14 *Caragana* genomes, with LTR retrotransposon lineages such as Tekay

contributing substantially to their genomic abundance (Fig. 3A). We further examined the relationship between genome size and the genomic abundance of major LTR retrotransposon families across the analyzed *Caragana* species (Additional file 1: Fig. S16). The genomic abundance of the Reina family was positively correlated with genome size (r = 0.75). Likewise, the relatively high genomic abundance of Ogre, Reina, Retand, and Tekay elements in XJCZ06 was associated with the comparatively large genome size of this species.

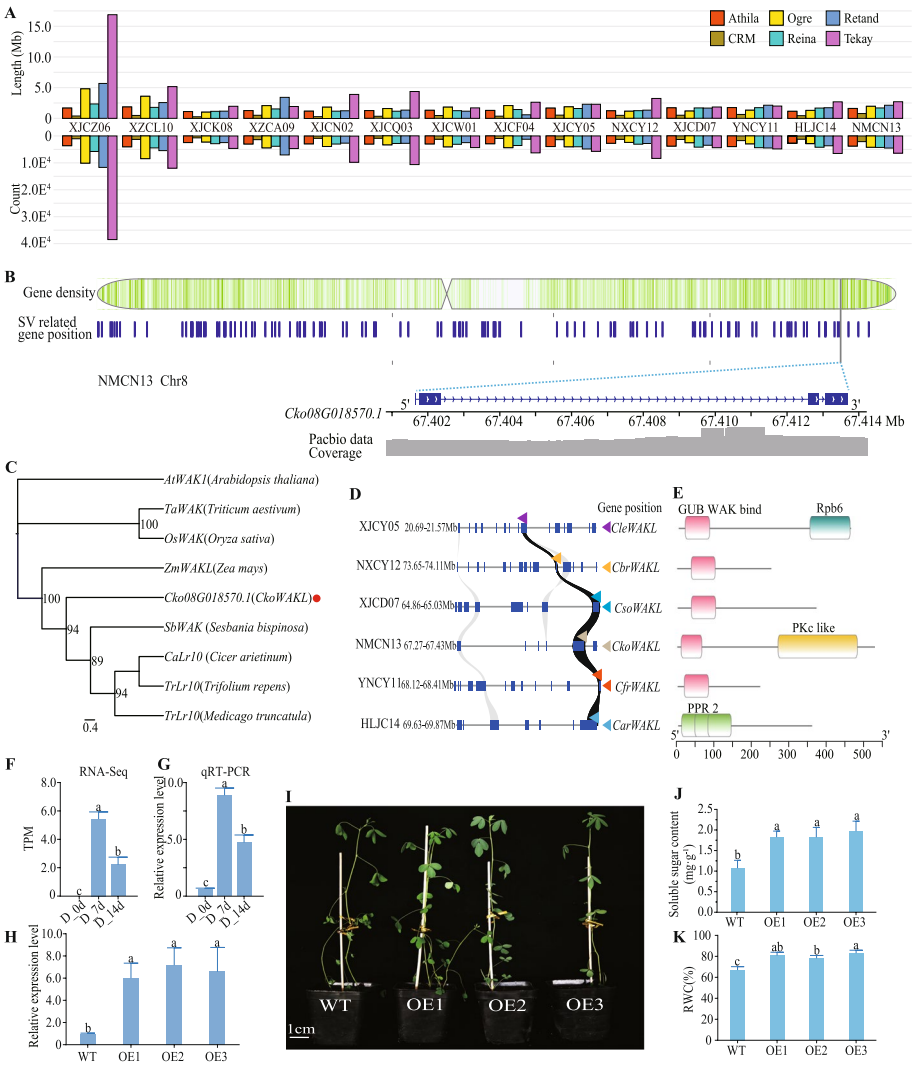


Fig. 3 LTR retrotransposon landscapes and functional characterization of *CkoWAKL*. **A** The size of the LTR subfamily in each *Caragana* species and the size relative to the LTR-RT subfamily. **B** Gene density, relative position of SVs and the position of *CkoWAKL* in Chr8 of NMCN13 (Values were calculated in 100-kb windows). The gray line group indicates the genome coverage when PacBio reads were mapped to the NMCN13 reference genome. **C** Phylogenetic tree of *WAK/WAKL* genes in different species. **D** Chromosome location and collinearity of *WAKL* genes in *Caragana*. **E** Protein domain structure of the *WAKL* homologue in *Caragana*. **F** TPM (transcripts per million) of *CkoWAKL* gene in NMCN13 under drought stress. **G** The qRT-PCR of *CkoWAKL* gene in NMCN13 under drought stress. **H** Relative expression of *CkoWAKL* in transgenic alfalfa overexpression lines. **I** Phenotypes of wild-type (WT) and three independent *CkoWAKL* overexpression lines (OE-1, OE-2, and OE-3) under drought treatment (Drought-7d, seven days of water withholding). **J** and **K** Soluble sugar content and relative water content (RWC) of WT and overexpression lines after drought treatment

We next compared the distribution and functional enrichment of TE-overlapping genes between the analyzed genomes. To compare the distribution of TE-overlapping genes between the cultivated NMCN13 genome and the wild XJCZ06 genome, we conducted GO enrichment analysis on TE-overlapping genes (Additional file 1: Fig. S17). The results indicated that the TE-overlapping genes of XJCZ06 were mainly enriched in biological processes related to chromatin organization and nucleobase-containing small molecule metabolism (Additional file 1: Fig. S18a; Additional file 2: Table S22). These results indicate that TE-overlapping genes in XJCZ06 are enriched in functional categories related to chromatin organization and small-molecule metabolism. In NMCN13, TE-overlapping genes were mainly enriched in processes such as nucleotidyltransferase activity and peptidase activity (Additional file 1: Fig. S18b; Additional file 2: Table S23). These findings highlight the unique functional enrichment patterns of different species, likely reflecting their distinct roles in regulating key cellular processes and their potential contributions to adaptation mechanisms. Most *Caragana* species showed a peak of relatively young intact LTR-RTs, whereas XJCW01, XJCF04, and XJCY05 displayed broader age distributions with a greater proportion of older intact elements (Additional file 1: Fig. S19).

CkoWAKL enhances drought-response phenotypes in transgenic alfalfa

Plant receptor kinases (PRKs) are integral to regulating various physiological processes in plants, including growth, development, and responses to environmental stress [41, 42]. From our whole genome analysis, we identified approximately 16,121 protein kinase genes across all the 14 *Caragana* species that we studied (Additional file 2: Table S24). These genes predominantly belong to the core gene family. Compared to other species, a significant proportion (90.49%) of genes in NMCN13 are from the core gene family, which may associate them with environmental adaptability traits (Additional file 2: Table S24). The SV analysis of the NMCN13 genome revealed an SV-associated gene, *Cko08G018570.1*, on Chr8 (Fig. 3B). Phylogenetic analysis showed that this gene belongs to the WAK gene family and has high homology with *ZmWAKL* [43] (Fig. 3C). In our synteny analysis of the 14 genomes, we observed that *WAK* maintained strong synteny in six of the genomes, namely XJCY05, XJCD07, YNCY11, NXCY12, NMCN13, and HLJC14 (Fig. 3D). These six species belong to clade III and have high genetic similarity (Fig. 1C).

In addition, we analyzed the protein domains of these six homologous genes and found that they contained four types of domains: GUB-WAK-bind (wall-associated receptor kinase galacturonan-binding), RPB6 (DNA-directed RNA Polymerase Subunit 6), PKC-like (protein kinase-like catalytic domain) and PPR2 (Pentatricopeptide Repeat 2) (Fig. 3E). The *CkoWAKL* gene contains two domains: GUB-WAK-bind and PKC-like. In contrast, the genes *CbrWAKL*, *CsoWAKL*, and *CfrWAKL* each contain only a single GUB-WAK-bind domain. The *CleWAKL* gene also possesses one GUB-WAK-bind along with an RPB6 domain, while the *CarWAKL* gene is distinct, containing three PPR2 domains (Fig. 3E). While *WAK/WAKL* genes have been predominantly implicated in plant disease resistance and immunity [43, 44], we observed that *CkoWAKL* was significantly upregulated in NMCN13 when plants were subjected to 7 and 14 days of drought stress (Fig. 3F and G; Additional file 1: Fig. S20; Additional file 2: Table S25). We then

overexpressed *CkoWAKL* in alfalfa (*Medicago sativa*) and obtained three independent overexpression lines (Fig. 3H). After seven days of water-withholding treatment, the overexpression lines displayed less severe wilting than the wild-type control (Fig. 3I). More importantly, the transgenic lines showed significantly higher relative water content (RWC) and soluble sugar content than the control plants under drought treatment (Fig. 3J and K). These results support the inference that *CkoWAKL* can enhance drought-response phenotypes and associated physiological indices in a heterologous system.

Discussion

The genus *Caragana* comprises perennial leguminous shrubs that are widely distributed in arid regions, where they display strong environmental adaptability and biological nitrogen fixation capabilities [45–47]. The long life cycle of *Caragana*, which is approximately 3–4 years from the time of sowing to seed harvest, complicates genetic breeding efforts [48]. A high-quality genome assembly of *Caragana* is crucial for understanding the genetic basis of its ecological adaptability and provides a valuable resource for further investigations into its physiology, biochemistry, and responses to environmental stresses. By conducting comparative genomic analyses and estimating divergence times, we found that the diversification of several *Caragana* lineages broadly coincided with the period of rapid uplift of the Qinghai-Tibet Plateau [29, 49]. The present-day occurrence of taxa such as XJCK08, XZCA09, and XZCL10 in the Qinghai-Tibet Plateau is consistent with this broader biogeographical context. Current biogeographical distribution patterns of the genus *Caragana* across Central Asia, East Asia, and the Qinghai-Tibet Plateau, are consistent with previous studies [50, 51]. Because root-nodule symbiosis is a defining feature of many legumes, we examined whether several well-established nodulation and symbiotic- signaling genes, including *NIN*, *DMI1*, *DMI2*, and *NSP1*, were retained across the 14 *Caragana* genomes. Homologues of these genes were detected in all sampled genomes and were assigned to core gene families, indicating that they are part of the conserved genomic repertoire rather than dispensable or lineage-restricted components of the *Caragana* pan-genome. In particular, *NIN* homologues retained the characteristic PB1 and RWP-RK domains of NIN-like proteins. Thus, our analysis supports conservation of major nodulation-related genes at the gene-family and domain-architecture levels. However, this does not imply that nodulation efficiency, rhizobial specificity, or symbiotic performance is identical among *Caragana* species, which will require direct physiological and microbiological testing.

Pan-genome research in plants is an emerging field in genomics. Whole-genome analyses can be leveraged to gain comprehensive insights into SVs that are closely associated with traits that are beneficial for crop breeding, biodiversity conservation, and evolutionary studies [22, 52]. In this study, we identified 53.5% core gene families, 4.2% private gene families and 42.3% dispensable gene families. We constructed a graph-based pan-genome with a total sequence span of 6.32 Gb, comprising 95,365,780 nodes and 122,449,332 edges, using NMCN13 as the linear reference. SVs are major sources of genetic variation and may influence adaptation in a changing environment [30, 53]. To date, most studies on *WAK/WAKL* genes have mainly focused on their effect on disease resistance [43, 44, 54, 55]. In addition to *CkoERF1*, we identified *CkoWAKL* as an SV-associated candidate gene and found that its expression was induced by drought stress

in NMCN13. Overexpression of *CkoWAKL* in alfalfa was associated with improved drought-response phenotypes, together with increased relative water content and soluble sugar content under short-term water-withholding treatment. These findings support a drought-response role for *CkoWAKL* in a heterologous system and illustrate the potential value of functional genomics for crop improvement.

In addition to directly affecting genetic diversity, SVs can affect the expression of nearby genes and guide the identification of genes associated with agronomically beneficial traits like drought stress tolerance [25, 30, 56]. PAVs have been strongly associated with traits that are relevant for crop improvement [27, 57]. The PAV-related genes we identified provide genetic evidence for understanding the environmental adaptability of *Caragana* species. The effects of PAVs on abiotic stress responses and associated regulatory mechanisms in the genus *Caragana* were largely unknown. To prioritize drought-related candidate genes from the pan-genome, we first examined PAV-related genes and identified *CkoERF1* as a candidate gene of interest. Separately, we identified *CkoWAKL* as a candidate because it was located within an SV hotspot in the NMCN13 genome and showed drought-responsive expression. We then evaluated the potential functional significance of these two genes using heterologous overexpression assays. These heterologous overexpression assays provide supportive functional evidence that these genes can influence drought-response phenotypes under experimental conditions (Figs. 2H and 3I). This study also provides valuable genome-wide resources for *Caragana* species to aid further investigative work on ecological genomics and population genetics.

Conclusions

In summary, this study provides high-quality chromosome-level genome assemblies for 14 ecologically diverse *Caragana* species, including a near-T2T assembly of *C. korshinskii*, and constructs a comprehensive graph-based pan-genome that captures extensive structural and presence-absence variations. Additionally, this work identifies *CkoERF1* and *CkoWAKL* as drought-related candidate genes whose overexpression enhances drought-response phenotypes in heterologous systems. Beyond these, we delineate phylogenetic relationships among the species, reveal clade-restricted gene-family expansions in pathways potentially related to stress responses, characterize transposable element dynamics associated with genome evolution, and identify conserved symbiosis-related core genes across the genus. Finally, these findings provide a comparative framework for investigating the environmental resilience in the genus *Caragana*, offering valuable resources for advancing functional genomics, evolutionary studies, and breeding programs aimed at enhancing drought tolerance in leguminous shrubs.

Methods

Plant materials and sequencing

The 14 *Caragana* species originated from diverse regions of China. Among them, two species, *Caragana polourensis* (XJCK08) and *Caragana arborescens* (HLJC14), were provided by the Kunming Institute of Botany, Chinese Academy of Sciences. The remaining 12 seed samples *Caragana acanthophylla* (XJCW01), *Caragana pumila* (XJCN02), *Caragana altaica* (XJCQ03), *Caragana tragacanthoides* (XJCF04), *Caragana leucospina* (XJCY05), *Caragana jubata* (XJCZ06), *Caragana soongorica* (XJCD07), *Caragana*

versicolor (XZCA09), *Caragana bicolor* (XZCL10), *Caragana pekinensis* (YNCY11), *Caragana roborovskyi* (NXCY12) and *Caragana korshinskii* (NMCN13) were collected from various locations across China, as depicted in Fig. 1A. These 14 species were selected to represent a broad portion of the natural distribution of the genus *Caragana* across temperate Asia, rather than only the collection sites of the focal accessions. *C. acanthophylla*, *C. pumila*, *C. altaica*, *C. tragacanthoides*, *C. leucospina*, and *C. soongorica* are centered in Xinjiang and adjacent Central Asia, where they typically occur in desert margins, dry steppe, rocky slopes, alluvial fans, and montane shrublands. *C. jubata*, *C. polourensis*, *C. versicolor*, and *C. bicolor* are characteristic of the Qinghai-Tibetan Plateau and neighboring high-elevation regions, where they occur in alpine shrublands, gravel slopes, river terraces, and cold semi-arid mountain grasslands. *C. pekinensis*, *C. roborovskyi*, *C. korshinskii*, and *C. arborescens* are distributed mainly in northern China and Inner Asia, occupying habitats such as loess hills, sandy or semi-fixed dunes, dry mountain slopes, and forest margins (Fig. 1A; Additional file 1: Fig. S2). Collectively, these species span arid, semi-arid, montane, and alpine environments across a broad east–west and altitudinal gradient, capturing much of the ecological and biogeographical breadth recognized for the genus *Caragana*. The 14 sampled *Caragana* species represent a broad geographical transect across the genus, ranging from northwestern China and adjacent Central Asia to the Qinghai-Tibetan Plateau and northern China (Additional file 1: Fig. S2), thereby covering much of the ecological and biogeographical breadth previously recognized for *Caragana* [11, 50, 58]. The collected seeds were screened and the husks of viable seeds were scratched for germination.

The germinated seeds were then transplanted into 10 cm × 10 cm × 20 cm pots filled with a mixture of vermiculite and quartz sand (1:1 ratio). Each pot contained a single plant, and each genotype was represented by 10 plants. During this period, the plants were irrigated with half-strength Hoagland nutrient solution every two days. All plant materials were grown in the cultivation chamber at the College of Grassland Agriculture, Northwest A&F University, under controlled conditions, the temperature was maintained 24 °C during the photoperiod (14 h of light) and 22 °C during the scotoperiod (10 h of darkness). At 13 weeks, leaves from each genotype were collected, immediately frozen in liquid nitrogen, and stored in −80 °C freezer for future use.

DNA extraction was performed using the Grandomics Genomic DNA Kit and ultra-long DNA was isolated using the BAC-long DNA Kit. The quality and quantity of total DNA were assessed using a NanoDrop One UV–Vis spectrophotometer (Thermo Fisher Scientific) and a Qubit 3.0 fluorometer (Invitrogen Life Technologies). Total RNA from the 14 samples was extracted using the RNAprep Pure Plant Kit (Tiangen, Beijing, China) according to the manufacturer's instructions. Library preparation for sequencing, as well as Illumina, Hi-C, PacBio, and ultra-long read sequencing (Oxford Nanopore) was conducted at Wuhan GrandOmics Biosciences Co., Ltd.

Genome assembly and evaluation

The genome size of the *Caragana* species was estimated using the findGSE package in R (parameters: sizek = 21, exp_hom = 140) [59]. PacBio sequencing reads were assembled into initial contigs using the Hifiasm software (v.0.13-r308) (–primary -l2 –h1 –h2) [60]. The Purge_haplotigs software (v.1.1.0) was employed to manage heterozygous regions

in the genome [61]. Initially, the read depth histogram was obtained using readhist and the thresholds for contigcov and purge were adjusted to eliminate redundant regions in the genome. The pairwise interactions among the non-redundant contigs were converted into interaction matrix files using 3D-DNA (v201008) and Juicebox(v1.11.08) [62] software (with default parameters), and manual sorting and orientation were performed using Juicebox. To obtain a complete genome, the preliminary assembly of the NMCN13 genome was gap filled using tgsgapcloser software (v1.2.1) (–min_nread 1 –min_match 2000 –minmap_arg '-x asm5') with the aid of ONT ultralong read data [63]. Three rounds of correction were performed using nextpolish (v1.4.1) to align Illumina data with the preliminary genome assembly, resulting in the final complete genome [64]. For 14 *Caragana* genomes, TIDK [65] was used to identify potential telomere repeat units and quarTeT (v 1.2.5) [66] was applied to search for potential telomere sequences in the genome using these repeat units, thus obtaining telomere location information and sequences.

To assess the quality and completeness of the assembly, Minimap2 software (v2.26-r1175) was used to map PacBio sequencing reads to each genome (with default parameters) [67]. Subsequently, SAMtools (v1.14) was used to calculate the genome coverage [68]. Candidate LTR retrotransposon sequences were first identified in each genome using LTR_FINDER and LTRharvest. The resulting candidates were then integrated and filtered using LTR_retriever to obtain a set of high-confidence intact LTR retrotransposon candidates. The Long Terminal Repeat Assembly Index (LAI) was subsequently calculated from this filtered intact LTR dataset. Merqury [69] (v1.3) was used to check the consistency of the assembled genome and obtain the consistency quality value (QV). The integrity of the assembled gene regions was evaluated using BUSCO (v5.2.0) [70] and 1,614 conserved core plant genes (embryophyta_odb10).

Gene prediction and annotation

For gene prediction and annotation, we first employed RepeatModeler (v2.0.1) [71] with the LTRStruct option to construct a de novo repeat sequence database. This predicted database was then classified using known repeat databases from Repbase (v20.11) [72] and Dfam (v3.5) [73] for Viridiplantae. The combined de novo predictions and known databases were subsequently duplicated to obtain a species-specific repeat sequence database. Using this custom-built repeat sequence database, RepeatMasker (v4.1.2) [74] with the options -e rmblast -gff -a -qq was applied to annotate repetitive sequences in the genome. Annotation of gene structures involved three primary steps. Initially, transcriptome sequences were assembled using Trinity (v2.11) [75]. Subsequently, the assembled transcripts were used for gene prediction with PASA (v2.4.1) [76]. The protein sequences of *A. thaliana*, *M. truncatula*, and *G. soja* were mapped to the genome using miniprot (v0.13) [77] with the parameters –gff –outc 0.5 –outn 100, resulting in a genome annotation file. Following this, the PASA results were used to train gene models with Augustus (v3.1.0) [78] (–hintsfile = hints.gff –sample = 1000 –genemodel = complete –softmasking = 1 –codingseq = on –exonnames = on). The genome file, with

masked repetitive sequences, was then subjected to gene structure prediction. Finally, EVM (v1.1.1) [79] was utilized to integrate the predicted results from the three aforementioned methods, yielding gene structure files for each genome. These files were further annotated with detailed functional information using eggnog-mapper (v2.1.12) [80] with default parameters.

Genomic evolutionary analysis

In addition to the 14 species of the *Caragana* species, we selected 17 additional plants, including *Arabidopsis thaliana* [81], *Cercis chinensis* [82], *Senna tora* [83], *Lupinus angustifolius* [84], *Lupinus albus* [85], *Cajanus cajan* [86], *Phaseolus vulgaris* [87], *Vigna radiata* [88], *Vigna angularis* [89], *Pueraria montana* [90], *Glycine soja* [91], *Pisum sativum* [92], *Medicago sativa* [93], *Medicago truncatula* [94], *Melilotus officinalis* [95], *Cicer arietinum* [96], and *Oxytropis ochrocephala* [97], for a comprehensive genomic evolutionary analysis. The genomes of 31 species (14 *Caragana* species, 16 Fabaceae species, and *Arabidopsis thaliana*) were analyzed using OrthoFinder (v2.5.5) [98], and the genes of these species were clustered (-S diamond -M msa -T raxml). A species phylogenetic tree was constructed using RAxML (v8.2.13) [99] software, applying the parameters (-m PROTGAMMAJTT -f a -# 1000). The divergence time of these species was obtained from the literature [93], and the divergence time between *M. sativa* and *M. truncatula* was approximately 8.0 MYA. These data were then used to estimate divergence times for each species with the mcmctree function of the paml (v4.10.7) [100]. To further explore syntenic blocks across the 14 *Caragana* genomes, MCScanX was implemented. Finally, WGDI (v0.6.5) [101] was utilized to calculate the Ks distribution, providing insights into the Whole-Genome Duplication (WGD) events within the *Caragana*.

Pan-genome construction and structural variant (SV) identification

We constructed a pan-genome based on 14 *Caragana* genomes. OrthoFinder (v2.5.5) [98] was used to estimate the gene sets in these 14 *Caragana* genomes by clustering of gene families, classifying them into three categories: core (genes present in all 14 species), dispensable (genes present in 2–13 species), and private (genes present only in one species). To identify dispensable gene families associated with environmental factors in the genomes of 14 *Caragana* species, we analyzed gene family presence/absence patterns derived from clustering analysis using OrthoFinder. Environmental data were downloaded from the National Earth System Science Data Center, National Science & Technology Infrastructure of China (http://www.geodata.cn). For each of the 14 sampled *Caragana* accessions, temperature, wind speed, relative humidity, precipitation, radiation and elevation were extracted for the collection site of that accession. For climatic variables, annual values from 2010 to 2022 were averaged to obtain one representative site-level value for each accession. To explore whether dispensable gene-family presence/absence patterns varied with environmental conditions at the collection sites of the sampled accessions, we adapted the ordination strategy of Kang et al. (2023) [30]. A binary matrix of dispensable gene-family presence/absence across the 14 sampled *Caragana* accessions was Hellinger-transformed, subjected to redundancy analysis (RDA) using the rda() function in vegan [102]. The clusterProfiler package in R [103] was used to perform transcription factor analysis of dispensable genes.

To construct a genetic variation map for the 14 *Caragana* genomes, we aligned the other 13 genomes with the NMCN13 reference genome using MUMmer (v.4.0.0) [104]. NUCmer (-c 1000 -l 40) was used for genome alignment, and delta-filter was used to filter the alignments in one-to-one mode (−1 -i 90 -l 100). The show-coords function of MUMmer was utilized to align the data positions (-THrd). Syri (v.1.6.3) [105] was then applied to perform a syntenic analysis of the *Caragana* genomes, with the aim of identifying structural rearrangements (absence, duplications, translocation, inversion, and presence), and unaligned regions. The resulting variations were visualized using plotsr (v1.1.0) [106]. In the SVMU result files, insertions or deletions (labeled INS or DEL) were considered presence/absence variations (PAVs). The final PAV regions were obtained by filtering out sequences with a coverage of > 80%. Genes overlapping with PAV regions by > 80% were considered PAV-related genes. Variant detection results were annotated using the annovar software (https://annovar.openbioinformatics.org/). The insertions and deletions detected by PAV and SyRI were de-redundant and integrated into the NMCN13 linear reference genome using the vg tool (v.1.25.0) [107]. We used the PGGB [108] (v.0.5.4) tool to construct the graph pan-genome.

Repetitive element identification and classification in *Caragana* spp.

First, 14 *Caragana* genomes were subjected to de novo repeat identification. The species-specific transposable element (TE) libraries were constructed based on the results obtained using RepeatModeler (v2.0.1) [71], LTR_FINDER (v1.0.7) [109], LTR_retriever (v2.8) and LTRharvest (v1.5.9) [110]. The panEDTA was used to merge all individual TE libraries, constructing a pan-TE library [111]. The library was then masked using RepeatMasker (v4.1.2) [74] with default parameters. Full-length LTR retrotransposons were accurately identified using LTR_retriever to ensure that complete LTR retrotransposons were correctly identified. Finally, all results were further classified using TEsorter (v1.4.7) [112] (-db rexdb-plant) and referenced to known repetitive sequence databases (for Viridiplantae) from Repbase (v20.11) and Dfam (v3.5). TE-overlapping genes were identified by intersecting annotated gene models with genome-wide TE annotations in each genome. This category refers to genes overlapping annotated TE regions and is not restricted to canonical TE-encoded genes such as Gag, Pol, integrase, reverse transcriptase, or Env-like genes.

Drought stress treatment and response analysis

The NMCN13 seedlings of uniform growth at 13 weeks were selected for stress treatment. Before drought treatment, all cultivated seedlings were irrigated with sufficient water (with water dripping from the bottom of the culture pot as the standard) and then weighed. Subsequently, seedlings were divided into three groups with three replicates of each: a control group (0d) where water was replenished every two days to maintain weight before drought treatment, and two treatment groups where samples were collected after continuous drought treatment until the weight of the entire pot remained unchanged, followed by additional drought periods of 7 and 14 days. Leaf and stem samples from the treatment groups were used for RNA extraction, promptly frozen in liquid nitrogen, and stored in a −80 °C freezer. Fresh samples were also measured relative water content [113] and malondialdehyde (MDA) content. The MDA content of

leaves was measured using a MDA Content Assay Kit (Solarbio, Beijing) according to the instructions. The raw data were processed using fastp to identify and remove low-quality sequences with default parameters. The clean reads were aligned to the NMCN13 genome using HISAT2 (v2.1.0) [114]. Gene expression was normalized as TPM (transcripts per million) using featureCounts with default parameters [115].

Evolution analysis of *DMI1* and *WAK/WAKL* genes

A global alignment was performed based on the *M. truncatula* DMI1 (Does Not Make Infections 1), NIN (Nodule Inception), NSP1 (Nodulation Signaling Pathway 1), and DMI2 (Does Not Make Infections 2) protein sequence and the *Caragana* protein sequence. The MtDMI1, MtNIN, MtNSP1, and MtDMI2 protein sequence was downloaded from NCBI. To identify genes encoding the AP2/ERF domain (Pfam: PF00847) in the 14 *Caragana* genomes, we performed a genome-wide search using reference *AP2/ERF* gene sets from *M. truncatula* and *A. thaliana* (obtained from PlantTFDB). Protein sequences from the 14 *Caragana* genomes were searched against the PF00847 hidden Markov model (HMM) profile using HMMER (v3.1) [116]. Additionally, an all-to-all BLASTP search was conducted using *M. truncatula* and *A. thaliana* AP2/ERF proteins as queries to identify candidate AP2/ERF genes in *Caragana*. All *Caragana* protein sequences were searched for members of protein kinase family using the HMM (PF00069) model downloaded from the Pfam database (http://pfam.xfam.org/). The sequence information of *AtWAK1*, *TaWAK*, *OsWAK* and *ZmWAKL* were downloaded from NCBI. The protein domains were predicted using the Conserved Domain Database (CDD) of NCBI. Multiple sequence alignment was performed using Muscle (v3.8) [117]. IQ-TREE (v2.3.6) [118] software was further applied to build the maximum likelihood (ML) phylogenetic tree with poisson-corrected model (-m MFP -bb 1000).

Validation of transgenic plants

Total RNA was extracted from NMCN13 leaves using the Eastep® Super Total RNA Extraction Kit (Promega, Beijing, China) following the manufacturer's protocol. First-strand cDNA was synthesized with the HiScript® II 1st Strand cDNA Synthesis Kit (Vazyme, Nanjing, China). For the transformation of *A. thaliana*, the floral dip method with Col-0 ecotype was used [119]. For this purpose, the coding sequence (CDS) of *CkoERF1* was cloned from the cDNA of NMCN13 leaves and subcloned into a pHB vector under the control of the 35S promoter. The resulting pHB-*CkoERF1* construct was further transformed into *Agrobacterium tumefaciens* strain GV3101. Positive transgenic plants were selected for hygromycin resistance, and at least nine independent transgenic lines were obtained from the T3 generation for further analysis. For identified lines, seeds from wild-type (WT) and overexpression lines (OE1, OE2, OE3) were germinated on half-strength MS media for 5 days. Seedlings with similar growth were then transferred to pots containing vermiculite. After three weeks of cultivation under normal conditions, watering was stopped and the phenotypes were observed and photographed after 7 days of drought stress.

The verified *CkoWAKL* sequence was cloned into the dicotyledonous expression vector pCAMBIA3300, which contains the CaMV 35S promoter, for overexpression in alfalfa (Biorun, China). Finally, the pCAMBIA3300- *CkoWAKL* plasmid was introduced

into the *A. tumefaciens* strain EHA105. For alfalfa transformation, the leaf disk method was employed. Specifically, leaves of alfalfa (cv. Zhongmu No. 1) were used as explants and subjected to ultrasound treatment at 40 kHz for 8 min in an *A. tumefaciens* suspension. The infected explants were subsequently cultured in co-culture, selection, and regeneration media to produce regenerated transgenic plants. According to the above drought stress treatment, the transgenic alfalfa was subjected to drought stress for 7 days. After 7 days of drought stress, the phenotype was observed and the relative water content and soluble sugar content were measured. The soluble sugar content of leaves was measured using a Plant Soluble Sugar Content Assay Kit (Solarbio, Beijing) according to the instructions. Three independent transgenic lines were analyzed for each experiment, with data presented as mean ± SD ($n = 9$). Significance was determined by one-way ANOVA followed by Tukey's test ($P < 0.05$).

qRT-PCR

Total RNA was extracted from fresh young leaves of NMCN13 and transgenic alfalfa, as well as from wild-type (WT) plants, using TRIzol reagent (Tiangen, China) following the manufacturer's protocol. First-strand cDNA synthesis was performed using 2 μL of RNA as a template with the PrimeScript™ II 1 st Strand cDNA Synthesis Kit (Takara, China), according to the manufacturer's instructions. For qRT-PCR, the reaction was carried out on a LightCycler® 480 System (Roche, Switzerland) using the TB Green® Premix Ex Taq™ (Takara, China). Each reaction contained 10 μL of TB Green® Premix Ex Taq™, 1 μL of cDNA, 0.5 μL of each forward and reverse primer, and 8 μL of nuclease-free water. The primers used in this study are listed in Additional file 2: Table S26. Gene expression levels were calculated using the $2^{\wedge -\Delta\Delta Ct}$ method [120], with *Caragana* internal reference genes (*CiEF1α*: KC679842) used for normalization [121].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-026-04230-8.

Additional file 1: Supplementary figures and figure legends (Fig. S1–S20). Fig. S1. Environmental adaptability and application potential of *C. korshinskii*. Fig. S2. The geographic-ecological diversity of sampled *Caragana* accessions. Fig. S3. Genome size and heterozygosity of 14 *Caragana* species estimated using 21 K-mer frequency distributions. Fig. S4. Hi-C chromosome interaction heatmaps for the 14 *Caragana* genome assemblies. Fig. S5. Distribution of the canonical telomeric repeat sequence, predicted centromeric regions, and predicted 45S precursor rRNA loci in the assembled *Caragana* genomes. Fig. S6. Phylogenetic tree and divergence time estimates for 14 *Caragana* species and other Fabaceae. Fig. S7. Functional-domain enrichment of genes from expanded gene families in 14 *Caragana* species. Fig. S8. KEGG pathway enrichment analysis of the expanded genes of 14 *Caragana* species. Fig. S9. Distribution of synonymous substitution per synonymous sitevalues. Fig. S10. Homologous gene plots of different species. Fig. S11. Enrichment analysis of KEGG pathways for core, dispensable, and private genes. Fig. S12. Phylogenetic relationships of symbiosis-related genes in 14 *Caragana* species. Fig. S13. Enrichment of transcription factor families, protein domains, and stress-related functional categories among PAV-related genes. Fig. S14. Association between dispensable gene-family composition and environmental variables, and transcription factor annotation of dispensable genes in selected *Caragana* genomes. Fig. S15. Evolutionary and sequence analysis of AP2/ERF gene family in *Arabidopsis thaliana*, *Medicago truncatula* and 14 *Caragana* species. Fig. S16. Correlation between genome size and the cumulative genomic sequence assigned to major LTR retrotransposon families in *Caragana*. Fig. S17. Distribution of TE-overlapping genes in *C. jubata*and *C. korshinskii*. Fig. S18. GO annotation analysis of TE-overlapping genes. Fig. S19. Estimated age distributions of intact LTR retrotransposons in *Caragana*. Fig. S20. Physiological indicators under drought stress conditions.

Additional file 2: Supplementary tables and table notes (Table S1–S26). Table S1. Summary of PacBio HiFi and Oxford Nanopore ultra-long sequencing data used for the *C. korshinskii*genome assembly. Table S2. Summary of Hi-C and Illumina paired-end sequencing data used for the *C. korshinskii*genome assembly and polishing. Table S3. Coordinates and lengths of telomeric repeat regions in the *C. korshinskii*genome assembly. Table S4. Predicted centromere coordinates in the *C. korshinskii*genome assembly. Table S5. Summary of the sampled *Caragana* materials used in this study. Table S6. Summary of PacBio HiFi sequencing data for 13 additional *Caragana* accessions. Table S7. Summary

of Hi-C sequencing data for the 14 *Caragana* accessions. Table S8. Summary of Illumina paired-end whole-genome sequencing data for the 14 *Caragana* accessions. Table S9. Summary of genome survey and primary assembly statistics for the 14 *Caragana* genomes. Table S10. Statistics of genome assembled into chromosomes. Table S11. Evaluation of assembly completeness. Table S12. Summary of RNA-seq data used for gene annotation of the 14 *Caragana* genomes. Table S13. Summary of genes predicted across 14 genomes. Table S14. Clustering statistics of gene families. Table S15. Statistics table of 14 *Caragana* pan-genome results. Table S16. KEGG pathway enrichment results for core, dispensable, and private gene families. Table S17. Summary of nodulation-related homologous genes in 14 *Caragana* genomes. Table S18. Summary of genetic variations generated by comparing 13 genomes to the *C. korshinskii*. Table S19. Summary of PAV-related genes in 13 genomes. Table S20. Historical variation of five climate factors across 14 sampling sitesover a 13-year period. Table S21. Numbers of AP2/ERF genes in representative plant species and the 14 *Caragana* genomes. Table S22. GO annotation analysis of TE-overlapping genes in *C. jubata*. Table S23. GO annotation analysis of TE-overlapping genes in *C. korshinskii*. Table S24. Numbers of PF00069-containing protein kinase genes in the 14 *Caragana* genomes. Table S25. Summary of RNA-seq quality metrics in leaves and stems of *C. korshinskii*under drought stress. Table S26. Primers used in this study.

Acknowledgements
We thank the Kunming Institute of Botany, Chinese Academy of Sciences (CAS), for the provision of Caragana germplasm resources. We thank Professor Wen Wang and Dr. Hang Zong from Northwestern Polytechnical University for their constructive suggestions. We thank Professor Linkai Huang and Haidong Yan from Sichuan Agricultural University for their constructive suggestions on data analysis.

Peer review information
Wenjing She was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
P.Y., J.S., and J.A. designed the project. P.Y. and Q.L. supervised this project. J.S., H.X., W.L., J.H. and Q.L. collected samples and genome sequencing. X.S., Y.Z., H.C. and T.H. gave advice and edited the manuscript. J.A., H.X., W.L. and C.L. performed functional validation experiments. J.S., C.L., X.S. and J.H. analyzed the data and revised the manuscript. J.S. and J.A. drafted the manuscript, and all authors reviewed and approved the final version.

Funding
This work was supported by China Agriculture Research System of MOF and MARA, the Key Research and Development Program of the Xinjiang Uygur Autonomous Region (2023B02031).

Data availability
The raw sequencing data generated for genome assembly, genome annotation, and transcriptomic analysis of the 14 Caragana species are available in the NCBI Sequence Read Archive under BioProject accession PRJNA1495756 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1495756) [122]. The chromosome-scale genome assemblies, corresponding gene annotation files, and associated pan-genome resources generated in this study are available from figshare (https://doi.org/10.6084/m9.figshare.28426868) [123]. The source code and command-line workflows used for genome assembly, pan-genome construction, structural variation detection, and downstream analyses are publicly available from Zenodo (https://doi.org/10.5281/zenodo.19812339) [124]. The corresponding project record and associated sequencing resources are also available from the National Genomics Data Center, China National Center for Bioinformation, under BioProject accession PRJCA030077 (https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA030077) [125].

Declarations

Ethics approval and consent to participate
It is not applicable.

Competing interests
The authors declare no competing interests.

Received: 14 February 2026 Accepted: 29 July 2026
Published online: 01 August 2026

References

1. Lewis GP, Schrire B, Mackinder B, Lock M. Legumes of the world. Kew, UK: Kew Publishing; 2005.
2. Zander P, Amjath-Babu TS, Preissel S, Reckling M, Bues A, Schläfke N, et al. Grain legume decline and potential recovery in European agriculture: a review. Agron Sustain Dev. 2016;36:26.
3. Barton L, Thamo T, Engelbrecht D, Biswas WK. Does growing grain legumes or applying lime cost effectively lower greenhouse gas emissions from wheat production in a semi-arid climate. J Clean Prod. 2014;83:194–203.
4. Chiang F, Mazdiyasni O, AghaKouchak A. Evidence of anthropogenic impacts on global drought frequency, duration, and intensity. Nat Commun. 2021;12:2754.
5. Ault TR. On the essentials of drought in a changing climate. Science. 2020;368:256–60.

6. Liu W, Zhu J, Jia Q, Zheng X, Li J, Lou X, et al. Carbon sequestration effects of shrublands in Three-North Shelterbelt Forest region, China. Chin Geogr Sci. 2014;24:444–53.
7. Zhao Y, Wang L, Knighton J, Evaristo J, Wassen M. Contrasting adaptive strategies by *Caragana korshinskii* and *Salix psammophila* in a semiarid revegetated ecosystem. Agric For Meteorol. 2021;300:108323.
8. Wang K, Zheng J, Li G, Ma Z, Zhang X, Zhang Q, et al. *Caragana korshinskii* of different planting years altered the soil water, carbon, and nitrogen storage in the different layers within level ditch on slope in a semi-arid area. CATENA. 2024;242:108126.
9. Guan SH, Chen WF, Wang ET, Lu YL, Yan XR, Zhang XX, et al. *Mesorhizobium caraganae* sp. nov., a novel rhizobial species nodulated with *Caragana* spp. in China. Int J Syst Evol Microbiol. 2008;58:2646–53.
10. Yuan X, Li H, Yu X, Ji Z. Genetic adaptation of *Mesorhizobium* symbionts associated with *Caragana* in Northern China Deserts. Ecol Evol. 2026;16:e73134.
11. Cui X, Li E, He J, Wang Y, Shang C, Zhong B, et al. Ancient hybridization drives arid adaptation and species diversification in *Caragana* (Fabaceae). New Phytol. 2025;247:2454–72.
12. Mei F, Yang T, Chao H, Ma X, Wu J, Yang Q, et al. Genomic insights into drought adaptation of the forage shrub *Caragana korshinskii* (Fabaceae) widely planted in drylands. Plant J. 2025;121:e17255.
13. Yang W, Feng H, Zhang X, Zhang J, Doonan JH, Batchelor WD, et al. Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. Mol Plant. 2020;13:187–214.
14. Burghardt LT. Evolving together, evolving apart: measuring the fitness of rhizobial bacteria in and out of symbiosis with leguminous plants. New Phytol. 2020;228:28–34.
15. Marks RA, Hotaling S, Frandsen PB, VanBuren R. Representation and participation across 20 years of plant genome sequencing. Nat Plants. 2021;7:1571–8.
16. Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. Plant pan-genomes are the new reference. Nat Plants. 2020;6:914–20.
17. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, et al. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet. 2011;43:956–63.
18. Sun Y, Shang L, Zhu Q-H, Fan L, Guo L. Twenty years of plant genome sequencing: achievements and challenges. Trends Plant Sci. 2022;27:391–401.
19. Ebler J, Ebert P, Clarke WE, Rausch T, Audano PA, Houwaart T, et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. Nat Genet. 2022;54:518–25.
20. Xia E, Tong W, Hou Y, An Y, Chen L, Wu Q, et al. The reference genome of tea plant and resequencing of 81 diverse accessions provide insights into its genome evolution and adaptation. Mol Plant. 2020;13:1013–26.
21. Qin P, Lu H, Du H, Wang H, Chen W, Chen Z, et al. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. Cell. 2021;184:3542-3558.e3516.
22. Gui S, Wei W, Jiang C, Luo J, Chen L, Wu S, et al. A pan-*Zea* genome map for enhancing maize improvement. Genome Biol. 2022;23:178.
23. Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, et al. Pan-genome of wild and cultivated soybeans. Cell. 2020;182:162-176.e113.
24. Jayakodi M, Padmarasu S, Haberer G, Bonthala VS, Gundlach H, Monat C, et al. The barley pan-genome reveals the hidden legacy of mutation breeding. Nature. 2020;588:284–9.
25. Shi T, Zhang X, Hou Y, Jia C, Dan X, Zhang Y, et al. The super-pangenome of *Populus* unveils genomic facets for its adaptation and diversification in widespread forest trees. Mol Plant. 2024;17:725–46.
26. Chen S, Wang P, Kong W, Chai K, Zhang S, Yu J, et al. Gene mining and genomics-assisted breeding empowered by the pangenome of tea plant *Camellia sinensis*. Nat Plants. 2023;9:1986–99.
27. Yan H, Sun M, Zhang Z, Jin Y, Zhang A, Lin C, et al. Pangenomic analysis identifies structural variation associated with heat tolerance in pearl millet. Nat Genet. 2023;55:507–18.
28. Ant Group. Gross Ecosystem Product (GEP) Accounting Report for the Ant Forest Afforestation Project, 2016–2020. Beijing, China: International Union for Conservation of Nature China Office; Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences. 2021. https://www.rcees.cas.cn/news/202308/W020230816532953776655.pdf.
29. Zhu W, Ding L, Ji Y, Qu R, Zhu Y, Xie C, et al. Subduction evolution controlled Himalayan orogenesis: implications from 3-D subduction modeling. Appl Sci. 2022;12:7413.
30. Kang M, Wu H, Liu H, Liu W, Zhu M, Han Y, et al. The pan-genome and local adaptation of *Arabidopsis thaliana*. Nat Commun. 2023;14:6259.
31. Sun Y, Fernie AR. Plant secondary metabolism in a fluctuating world: climate change perspectives. Trends Plant Sci. 2024;29:560–71.
32. Long W, He Q, Wang Y, Wang Y, Wang J, Yuan Z, et al. Genome evolution and diversity of wild and cultivated rice species. Nat Commun. 2024;15:9994.
33. Ané J-M, Kiss GB, Riely BK, Penmetsa RV, Oldroyd GED, Ayax C, et al. *Medicago truncatula DMI1* required for bacterial and fungal symbioses in legumes. Science. 2004;303:1364–7.
34. Schiessl K, Lilley JLS, Lee T, Tamvakis I, Kohlen W, Bailey PC, et al. NODULE INCEPTION recruits the lateral root developmental program for symbiotic nodule organogenesis in *Medicago truncatula*. Curr Biol. 2019;29:3657-3668.e3655.
35. Liu W, Kohlen W, Lillo A, Op den Camp R, Ivanov S, Hartog M, et al. Strigolactone biosynthesis in *Medicago truncatula* and rice requires the symbiotic GRAS-type transcription factors NSP1 and NSP2. Plant Cell. 2011;23:3853–65.
36. Pan H, Stonoha-Arther C, Wang D. *Medicago* plants control nodulation by regulating proteolysis of the receptor-like kinase DMI2. Plant Physiol. 2018;177:792–802.
37. Hu H, Scheben A, Wang J, Li F, Li C, Edwards D, et al. Unravelling inversions: technological advances, challenges, and potential impact on crop breeding. Plant Biotechnol J. 2024;22:544–54.
38. Feng K, Hou XL, Xing GM, Liu JX, Duan AQ, Xu ZS, et al. Advances in AP2/ERF super-family transcription factors in plant. Crit Rev Biotechnol. 2020;40:750–76.

39. Bennetzen JL, Wang H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu Rev Plant Biol. 2014;65:505–30.
40. Zhang S-J, Liu L, Yang R, Wang X. Genome size evolution mediated by gypsy retrotransposons in *Brassicaceae*. Genomics Proteomics Bioinformatics. 2020;18:321–32.
41. Liang X, Zhou J-M. Receptor-like cytoplasmic kinases: central players in plant receptor kinase–mediated signaling. Annu Rev Plant Biol. 2018;69:267–99.
42. Cai W, Hong J, Liu Z, Wang W, Zhang J, An G, et al. A receptor-like kinase controls the amplitude of secondary cell wall synthesis in rice. Curr Biol. 2023;33:498-506.e496.
43. Zhong T, Zhu M, Zhang Q, Zhang Y, Deng S, Guo C, et al. The ZmWAKL–ZmWIK–ZmBLK1–ZmRBOH4 module provides quantitative resistance to gray leaf spot in maize. Nat Genet. 2024;56:315–26.
44. Li Q, Hu A, Qi J, Dou W, Qin X, Zou X, et al. CsWAKL08, a pathogen-induced wall-associated receptor-like kinase in sweet orange, confers resistance to citrus bacterial canker via ROS control and JA signaling. Hortic Res. 2020;7:42.
45. Ma CC, Gao YB, Guo HY, Wang JL, Wu JB, Xu JS. Physiological adaptations of four dominant *Caragana* species in the desert region of the Inner Mongolia Plateau. J Arid Environ. 2008;72:247–54.
46. Xu F, Zhong XC, Sun RC, Lu Q. Anatomy, ultrastructure and lignin distribution in cell wall of *Caragana korshinskii*. Ind Crops Prod. 2006;24:186–93.
47. Zhang M, Xue J, Zhang Q, Sanderson SC. Inferring ancestral distribution area and survival vegetation of *Caragana* (Fabaceae) in Tertiary. Plant Syst Evol. 2015;301:1831–42.
48. Liu B, Shang X, Zhang X, Shao W, Ren L, Li G, Zhu M, Wang R. In vitro regeneration and Agrobacterium-mediated genetic transformation of Caragana korshinskii. For Res. 2023;3:3-14.
49. Miao Y, Fang X, Sun J, Xiao W, Yang Y, Wang X, et al. A new biologic paleoaltimetry indicating Late Miocene rapid uplift of northern Tibet Plateau. Science. 2022;378:1074–9.
50. Zhang M-L, Fritsch PW. Evolutionary response of *Caragana* (Fabaceae) to Qinghai–Tibetan Plateau uplift and Asian interior aridification. Plant Syst Evol. 2010;288:191–9.
51. Zhang M-L, Xiang X-G, Xue J-J, Sanderson SC, Fritsch PW. Himalayan uplift shaped biomes in Miocene temperate Asia: evidence from leguminous *Caragana*. Sci Rep. 2016;6:36528.
52. Schreiber M, Jayakodi M, Stein N, Mascher M. Plant pangenomes for crop improvement, biodiversity and evolution. Nat Rev Genet. 2024;25:563–77.
53. Song J-M, Guan Z, Hu J, Guo C, Yang Z, Wang S, et al. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. Nat Plants. 2020;6:34–45.
54. Zhang B, Su T, Xin X, Li P, Wang J, Wang W, et al. Wall-associated kinase BrWAK1 confers resistance to downy mildew in *Brassica rapa*. Plant Biotechnol J. 2023;21:2125–39.
55. Wang P, Zhou L, Jamieson P, Zhang L, Zhao Z, Babilonia K, et al. The cotton wall-associated kinase GhWAK7A mediates responses to fungal wilt pathogens by complexing with the chitin sensory receptors. Plant Cell. 2020;32:3978–4001.
56. Tao Y, Luo H, Xu J, Cruickshank A, Zhao X, Teng F, et al. Extensive variation within the pan-genome of cultivated and wild sorghum. Nat Plants. 2021;7:766–73.
57. He Q, Tang S, Zhi H, Chen J, Zhang J, Liang H, et al. A graph-based genome and pan-genome variation of the model plant *Setaria*. Nat Genet. 2023;55:1232–42.
58. Zhang ML. A dispersal and vicariance analysis of the genus *Caragana* Fabr. J Integr Plant Biol. 2005;47:897–904.
59. Sun H, Ding J, Piednoël M, Schneeberger K. FindGSE: estimating genome size variation within human and *Arabidopsis* using k-mer frequencies. Bioinformatics. 2017;34:550–7.
60. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18:170–5.
61. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19:460.
62. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 2016;3:99–101.
63. Xu M, Guo L, Gu S, Wang O, Zhang R, Peters BA, et al. TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. Gigascience. 2020;9:giaa094.
64. Hu J, Fan J, Sun Z, Liu S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics. 2020;36:2253–5.
65. Brown MR, Manuel Gonzalez de La Rosa P, Blaxter M. Tidk: a toolkit to rapidly identify telomeric repeats from genomic datasets. Bioinformatics. 2025;41:btaf049.
66. Lin Y, Ye C, Li X, Chen Q, Wu Y, Zhang F, et al. QuarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. Hortic Res. 2023;10:uhad127.
67. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
68. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
69. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020;21:245.
70. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.
71. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117:9451–7.
72. Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11.
73. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam database of repetitive DNA families. Nucleic Acids Res. 2016;44:D81-89.
74. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. 2009, Chapter 4:4.10.11–14.10.14.

75. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.
76. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003;31:5654–66.
77. Li H. Protein-to-genome alignment with miniprot. Bioinformatics. 2023;39:btad014.
78. Stanke M, Steinkamp R, Waack S, Morgenstern B. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 2004;32:W309-312.
79. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol. 2008;9:R7.
80. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. Mol Biol Evol. 2017;34:2115–22.
81. The *Arabidopsis* Genome I. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000;408:796–815.
82. Li J, Shen J, Wang R, Chen Y, Zhang T, Wang H, et al. The nearly complete assembly of the *Cercis chinensis* genome and Fabaceae phylogenomic studies provide insights into new gene evolution. Plant Commun. 2023;4:100422.
83. Kang S-H, Pandey RP, Lee C-M, Sim J-S, Jeong J-T, Choi B-S, et al. Genome-enabled discovery of anthraquinone biosynthesis in *Senna tora*. Nat Commun. 2020;11:5875.
84. Hane JK, Ming Y, Kamphuis LG, Nelson MN, Garg G, Atkins CA, et al. A comprehensive draft genome sequence for lupin (*Lupinus angustifolius*), an emerging health food: insights into plant-microbe interactions and legume evolution. Plant Biotechnol J. 2017;15:318–30.
85. Hufnagel B, Marques A, Soriano A, Marquès L, Divol F, Doumas P, et al. High-quality genome sequence of white lupin provides insight into soil exploration and seed quality. Nat Commun. 2020;11:492.
86. Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, et al. Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nat Biotechnol. 2012;30:83–9.
87. Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, et al. A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet. 2014;46:707–13.
88. Kang YJ, Kim SK, Kim MY, Lestari P, Kim KH, Ha B-K, et al. Genome sequence of mungbean and insights into evolution within *Vigna* species. Nat Commun. 2014;5:5443.
89. Kang YJ, Satyawan D, Shim S, Lee T, Lee J, Hwang WJ, et al. Draft genome sequence of adzuki bean, *Vigna angularis*. Sci Rep. 2015;5:8069.
90. Mo C, Wu Z, Shang X, Shi P, Wei M, Wang H, et al. Chromosome-level and graphic genomes provide insights into metabolism of bioactive metabolites and cold-adaption of *Pueraria lobata* var. *montana*. DNA Res. 2022;29:dsac030.
91. Li Y-h, Zhou G, Ma J, Jiang W, Jin L-g, Zhang Z, et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol. 2014;32:1045–52.
92. Kreplak J, Madoui M-A, Cápal P, Novák P, Labadie K, Aubert G, et al. A reference genome for pea provides insight into legume genome evolution. Nat Genet. 2019;51:1411–22.
93. Shen C, Du H, Chen Z, Lu H, Zhu F, Chen H, et al. The chromosome-level genome sequence of the autotetraploid alfalfa and resequencing of core germplasms provide genomic resources for alfalfa research. Mol Plant. 2020;13:1250–61.
94. Pecrix Y, Staton SE, Sallet E, Lelandais-Brière C, Moreau S, Carrère S, et al. Whole-genome landscape of *Medicago truncatula* symbiotic genes. Nat Plants. 2018;4:1017–25.
95. He Q, Li Z, Liu Y, Yang H, Liu L, Ren Y, et al. Chromosome-scale assembly and analysis of *Melilotus officinalis* genome for SSR development and nodulation genes analysis. Plant Genome. 2023;16:e20345.
96. Varshney RK, Song C, Saxena RK, Azam S, Yu S, Sharpe AG, et al. Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. Nat Biotechnol. 2013;31:240–6.
97. Zhang L, Wu R, Mur LAJ, Guo C, Zhao X, Meng H, et al. Assembly of high-quality genomes of the locoweed *Oxytropis ochrocephala* and its endophyte *Alternaria oxytropis* provides new evidence for their symbiotic relationship and swainsonine biosynthesis. Mol Ecol Resour. 2023;23:253–72.
98. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:157.
99. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.
100. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24:1586–91.
101. Sun P, Jiao B, Yang Y, Shan L, Li T, Li X, et al. WGDI: a user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. Mol Plant. 2022;15:1841–51.
102. Dixon P. VEGAN, a package of R functions for community ecology. J Veg Sci. 2003;14:927–30.
103. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. ClusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation (Camb). 2021;2:100141.
104. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: a fast and versatile genome alignment system. PLoS Comput Biol. 2018;14:e1005944.
105. Goel M, Sun H, Jiao W-B, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. 2019;20:277.
106. Goel M, Schneeberger K. Plotsr: visualizing structural similarities and rearrangements between multiple genomes. Bioinformatics. 2022;38:2922–6.
107. Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga J, et al. Genotyping structural variants in pangenome graphs using the vg toolkit. Genome Biol. 2020;21:35.
108. Garrison E, Guarracino A, Heumos S, Villani F, Bao Z, Tattini L, et al. Building pangenome graphs. Nat Methods. 2024;21:2008–12.
109. Ou S, Jiang N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. Mob DNA. 2019;10:48.

110. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics. 2008;9:18.
111. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019;20:275.
112. Zhang R-G, Li G-Y, Wang X-L, Dainat J, Wang Z-X, Ou S, et al. TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. Hortic Res. 2022;9:uhac017.
113. Turner NC. Techniques and experimental approaches for the measurement of plant water status. Plant Soil. 1981;58:339–66.
114. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15.
115. Liao Y, Smyth GK, Shi W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2013;30:923–30.
116. Prakash A, Jeffryes M, Bateman A, Finn RD. The HMMER web server for protein sequence similarity search. Curr Protoc Bioinformatics. 2017;60:3.15.11-13.15.23.
117. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
118. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37:1530–4.
119. Zhang X, Henriques R, Lin S-S, Niu Q-W, Chua N-H. *Agrobacterium*-mediated transformation of *Arabidopsis thaliana* using the floral dip method. Nat Protoc. 2006;1:641–6.
120. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods. 2001;25:402–8.
121. Yang Q, Yin J, Li G, Qi L-w, Yang F, Wang R, et al. Reference gene selection for qRT-PCR in *Caragana korshinskii* Kom. under different stress conditions. Mol Biol Rep. 2014;41:2325–34.
122. Song J, An J, Xia H, Li W, Hu J, Li Q, Li C, Shen X, Cheng H, Zhou Y, Hu T, Yang P. Raw sequencing data for genome assembly, genome annotation, and transcriptomic analysis of 14 *Caragana* species. Datasets. NCBI Sequence Read Archive. BioProject PRJNA1495756. 2026. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1495756.
123. Song J, An J, Xia H, Li W, Hu J, Li Q, Li C, Shen X, Cheng H, Zhou Y, Hu T, Yang P. Pan-genome resources for 14 *Caragana* species. figshare; 2026. https://doi.org/10.6084/m9.figshare.28426868.
124. Song J, An J, Xia H, Li W, Hu J, Li Q, et al. Code and workflows for pan-genome analysis of *Caragana* species. 2026. Zenodo. https://doi.org/10.5281/zenodo.19812339.
125. Song J, An J, Xia H, Li W, Hu J, Li Q, Li C, Shen X, Cheng H, Zhou Y, Hu T, Yang P. Raw sequencing resources for the *Caragana* pan-genome project. Datasets. National Genomics Data Center, China National Center for Bioinformation. BioProject PRJCA030077. 2026. https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA030077.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.