nature genetics

Article

https://doi.org/10.1038/s41588-025-02164-8

Pan-genomic analysis highlights genes associated with agronomic traits and enhances genomics-assisted breeding in alfalfa

Received: 21 February 2024

Accepted: 13 March 2025

Published online: 23 April 2025

Check for updates

Fei He[1,5], Shuai Chen[2,5], Yangyang Zhang[1,5], Kun Chai[2], Qing Zhang[3], Weilong Kong[2], Shenyang Qu[2], Lin Chen[1], Fan Zhang[1], Mingna Li[1], Xue Wang[1], Huigang Lv[1], Tiejun Zhang[4], Xiaofan He[4], Xiao Li[4], Yajing Li[1], Xianyang Li[1], Xueqian Jiang[1], Ming Xu[1], Bilig Sod[1], Junmei Kang[1], Xingtan Zhang[2] ✉, Ruicai Long[1] ✉ & Qingchuan Yang[1] ✉

Alfalfa (*Medicago sativa* L.), a globally important forage crop, is valued for its high nutritional quality and nitrogen-fixing capacity. Here, we present a high-quality pan-genome constructed from 24 diverse alfalfa accessions, encompassing a wide range of genetic backgrounds. This comprehensive analysis identified 433,765 structural variations and characterized 54,002 pan-gene families, highlighting the pivotal role of genomic diversity in alfalfa domestication and adaptation. Key structural variations associated with salt tolerance and quality traits were discovered, with functional analysis implicating genes such as *MsMAP65* and *MsGA3ox1*. Notably, overexpression of *MsGA3ox1* led to a reduced stem–leaf ratio and enhanced forage quality. The integration of genomic selection and marker-assisted breeding strategies improved genomic estimated breeding values across multiple traits, offering valuable genomic resources for advancing alfalfa breeding. These findings provide insights into the genetic basis of important agronomic traits and establish a solid foundation for future crop improvement.

Alfalfa is unrivaled among forage crops because of its unique combination of high quality, high yield, biological nitrogen fixation and stress resistance properties. Originally introduced as an historically important crop and undergoing extensive domestication, alfalfa has evolved into a high-yielding and adaptive forage crop after years of refinement and cultivation. As an essential perennial legume crop, alfalfa's crude protein content exceeds 20% during its initial flowering stage, exhibiting high digestibility and rich nutritional value[1]. The abundant genetic diversity of alfalfa enables it to maintain relatively stable growth even under various environmental pressures. In addition, alfalfa forms a symbiotic relationship with rhizobia, facilitating atmospheric nitrogen fixation[2,3]. This symbiosis not only reduces reliance on nitrogen fertilizers during the growth period, but also contributes to economic savings and environmental benefits[2,3]. Therefore, alfalfa plays a crucial role in the global food supply and soil remediation.

Although alfalfa has many potential advantages as a cross-pollinated autotetraploid plant ($2n = 4\times = 32$), compared with the diploid *Medicago truncatula* (Mt) ($2n = 2\times = 16$), its self-incompatibility and complex genetic characteristics pose challenges for adopting

[1]Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing, China. [2]National Key Laboratory for Tropical Crop Breeding, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China. [3]State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, College of Agriculture, Guangxi Key Laboratory of Sugarcane Biology, Guangxi University, Nanning, China. [4]School of Grassland Science, Beijing Forestry University, Beijing, China. [5]These authors contributed equally: Fei He, Shuai Chen, Yangyang Zhang. ✉e-mail: zhangxingtan@caas.cn; longruicai@caas.cn; yangqingchuan@caas.cn

https://doi.org/10.1038/s41588-025-02164-8

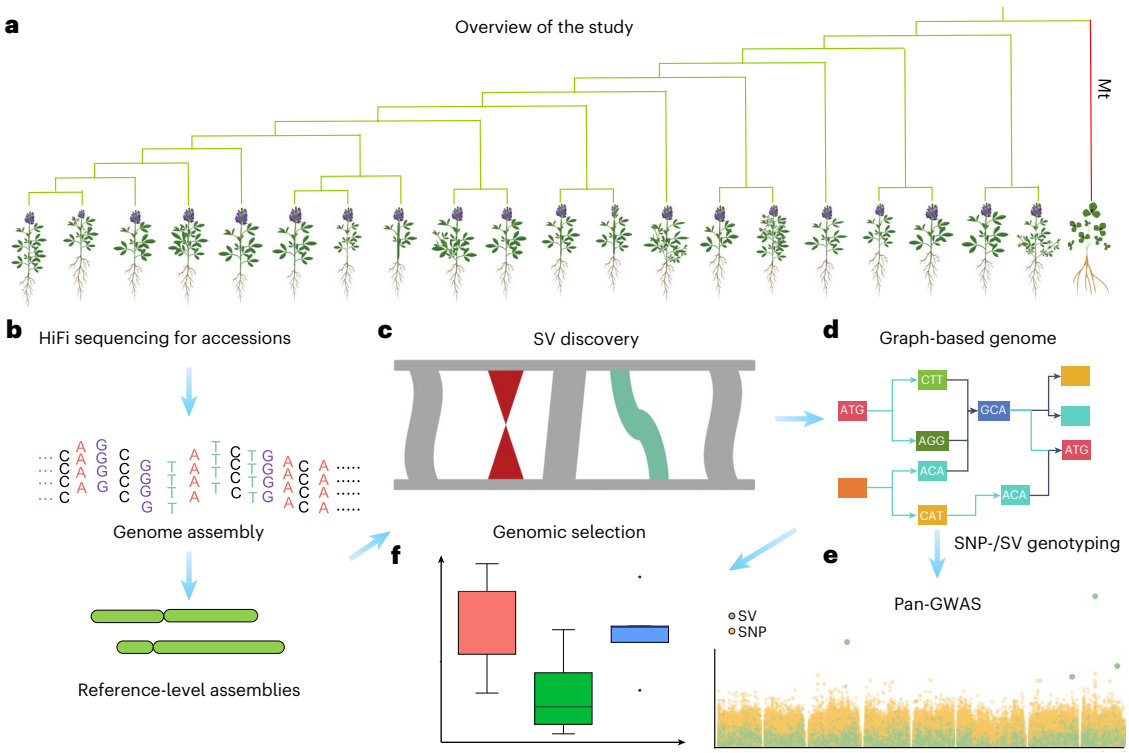


Fig. 1 | Layout of the alfalfa graph pan-genome study. a, Phylogenetic tree of diverse alfalfa cultivars selected based on phylogenetic relationships. Illustration created by the authors. No third-party materials were used. **b**, Assembled genomic sequence maps of the alfalfa cultivars. **c**, The multiple types of SVs identified. **d**, The pan-genome was constructed using the genome of cultivar ZM4 as a reference. **e**, Short-read sequences mapped onto the graph-based pan-genome to analyze SVs and SNPs associated with agronomic traits. The *x*-axis represents eight bars corresponding to chromosomes Chr1 to Chr8, and the *y*-axis indicates $-\log_{10}(p)$. **f**, Whole-genome selection analysis conducted to identify breeding potential in the alfalfa population. From left to right, the three bars represent Trait 1, Trait 2 and Trait 3, and the *y*-axis indicates improvement (%).

efficient breeding strategies. These complexities hinder the analysis and utilization of agronomic traits. Currently, three autopolyploid alfalfa genomes are available, XinJiangDaYe[4], Zhongmu No. 1 (ref. 2) and Zhongmu-4 (ZM4)[5]. These genomes provide essential tools for population genomics and molecular genetics research, contributing to a deeper understanding of the alfalfa domestication process and accelerating genetic improvements. By resequencing a large number of germplasm resources, these studies have thoroughly characterized the genetic diversity of alfalfa, highlighting its crucial role in improvement and laying a solid foundation for agronomic traits identification. However, to fully capture the complex genetic landscape of alfalfa, particularly its large structural variations (SVs), a shift toward more comprehensive genomic approaches is needed.

Conventional reliance on a single reference genome sequence is often insufficient to capture the full spectrum of species sequence diversity, particularly in the context of complex large SVs[6,7]. SVs are known to play a pivotal role in determining agronomic traits[8–11], crop domestication[12,13], genetic improvement[9] and evolutionary processes[14–16]. In this context, pan-genomes, which integrate complex variations from multiple genomes in a species, provide a more accurate and comprehensive approach to gene characterization. An emerging focus in genetic breeding research includes the construction of a comprehensive pan-genome that encompasses alfalfa's core germplasm, which can be used to analyze the genetic basis of key economic traits. This comprehensive approach is expected to drive the sustainable advancement and improvement of alfalfa, providing solid support for developing new varieties with higher yields and greater adaptability.

In this study, we selected diverse alfalfa cultivars based on phylogenetic relationships and constructed a phylogenetic tree (Fig. 1a). Subsequently, we assembled the genome sequences of these cultivars, thereby obtaining individual genomic maps (Fig. 1b). Through detailed analysis of these genomic datasets, we identified multiple types of SVs and illustrated their distribution across the genomes (Fig. 1c). Using the genome of cultivar ZM4 as a reference, we constructed a pan-genome that includes all accessions (Fig. 1d). To investigate the structural variants associated with agronomic traits, we mapped short-read sequences generated by Illumina technology onto this graph-based pan-genome (Fig. 1e). In particular, our research focused on alfalfa's salt stress response and quality traits. By combining genome-wide association studies (GWAS) and SV analysis, we identified key genetic markers associated with salt stress tolerance and quality traits. Finally, we performed a whole-genome selection analysis to assess the inherent breeding potential in the alfalfa population (Fig. 1f). These in-depth phenotypic observations enhance our understanding of the underlying genetic basis and provide valuable insights for accelerating crop domestication and improvement (Supplementary Fig. 1).

Results

Improvement of the monoploid genome assembly of ZM4 alfalfa

Given the complexity of the tetraploid genome, particularly the challenges in analyzing genetic variation and conducting functional genomics studies, we opted to use a monoploid genome. The monoploid genome of a polyploid organism simplifies genetic analysis by providing a complete representation of genetic diversity, while reducing complexity and size. This simplification facilitates downstream genetic and functional analyses. The initial assembled 2.74 Gb contigs of the ZM4 alfalfa genome were processed to remove redundant contigs using the Khaper[17] program, and the nonredundant contigs were then used to assemble the monoploid genome[5]. Following redundancy, we

https://doi.org/10.1038/s41588-025-02164-8

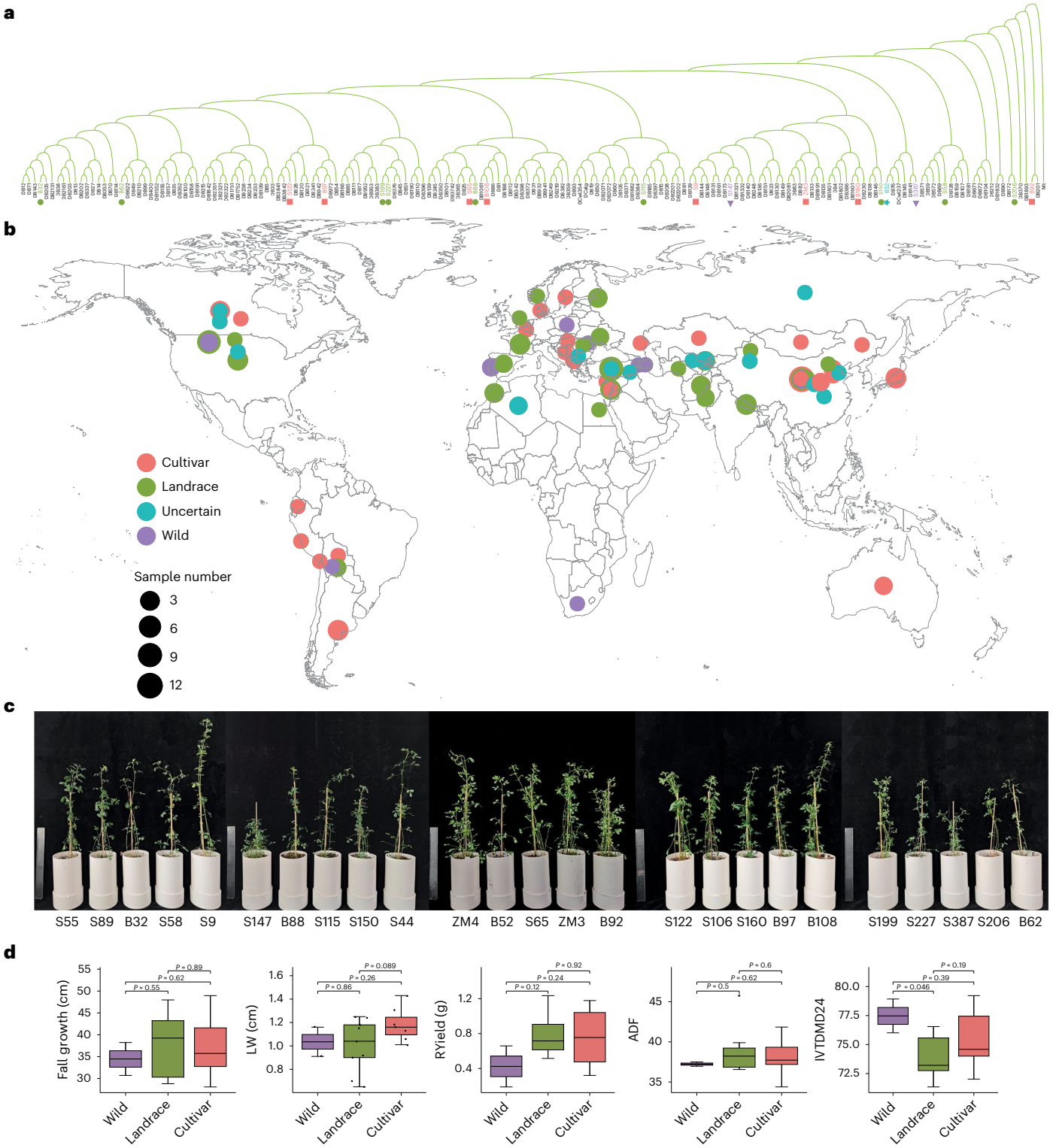


Fig. 2 | Distribution and diversity of representative alfalfa accessions. **a**, Phylogenetic analysis of 176 germplasm samples visualized using various colors and shapes to represent different germplasm categories. Pink squares denote cultivar accessions categories, green circles represent landrace accessions, cyan stars signify accessions of uncertain improvement status and purple triangles indicate wild accessions. **b**, Global distribution of the 176 diverse representative sample corresponding to the color scheme in Fig. 2a. Dot sizes depict sample quantity. The map was generated utilizing the R package ggplot2. No third-party materials were used. **c**, Comparative greenhouse growth profiles of diverse accessions (including the reference genome accession). **d**, Variations in traits such as fall growth, leaf width (LW), relative yield under salt stress (RYield), acid detergent fiber (AD) and in vitro for 24 h true dry matter degradability (IVTDMD24) across wild, landrace and cultivated categories, comprising two wild, nine landrace and eight cultivar accessions. In the boxplots, the 25th and 75th percentiles are represented by the bottom and top of the box, respectively, with the median shown as a line in the center. Outliers are displayed with whiskers extending to 1.5× the interquartile range. Statistical significance was calculated using two-sided Wilcoxon tests.

https://doi.org/10.1038/s41588-025-02164-8

Table 1 | Statistics of assemblies of 24 alfalfa accession genomes

Accession	Initial assembly (Mb)	Contig N50 (bp)	Monoploid assembly (Mb)	Contig N50 (bp)	Complete BUSCOs (C, %)	Complete and single-copy BUSCOs (S, %)	Complete and duplicated BUSCOs (D, %)	Fragmented BUSCOs (F, %)	Missing BUSCOs (M, %)	LAI	Repetitive sequences (%)
B32	3,686.5	512,873	854.6	1,087,817	90.7	77.2	13.5	0.5	8.8	13.57	61.11
B52	2,787	387,092	801.7	1,299,245	93.7	79.3	14.4	0.7	5.6	14.69	60.08
B62	3,594.5	474,292	837.8	990,738	95.1	83.5	11.6	0.7	4.2	14.52	57.33
B88	3,472.4	384,882	848.3	1,277,813	93	77.4	15.6	0.9	6.1	13.73	58.72
B92	3,366	576,057	861.4	1,406,841	92.6	77.9	14.7	0.7	6.7	15.02	59.16
B97	3,017.7	382,208	809.1	1,189,529	96.1	86.3	9.8	0.5	3.4	13.58	58.55
B106	3,342.2	491,397	831.3	1,361,605	93.3	78.4	14.9	0.7	6	14.83	57.68
B108	3,514	630,836	811.7	997,925	92.5	76.5	16	0.7	6.8	14.35	55.11
B160	3,580.2	470,411	864.8	1,230,746	93.2	77.4	15.8	0.7	6.1	15.04	61.96
S9	3,483.7	292,737	886.7	747,688	93	77.9	15.1	0.2	6.8	14.21	61.66
S44	2,709.4	441,069	880.9	1,004,912	94.6	78.8	15.8	0.2	5.2	13.89	58.12
S55	3,281.7	331,338	827.4	1,112,568	92.8	76.5	16.3	0.9	6.3	13.48	58.66
S58	3,746.2	433,227	854.5	1,106,812	92.1	80.9	11.2	0.7	7.2	14.96	63.20
S65	3,635.5	190,885	869.6	945,434	90.7	75.8	14.9	0.9	8.4	14.03	58.41
S89	3,341	414,485	858.5	901,653	92.5	76.5	16	1.2	6.3	14.78	60.39
S115	3,576.2	456,477	881.1	1,036,630	94.2	78.4	15.8	0.5	5.3	14.20	60.61
S122	3,444.7	320,453	830.8	686,689	93.3	79.1	14.2	0.5	6.2	13.83	60.55
S147	3,733.3	396,299	883	940,419	93.5	77.9	15.6	0.9	5.6	14.02	61.36
S150	3,619.8	431,545	881.7	1,062,517	92.3	75.3	17	0.7	7	13.75	60.99
S199	3,321.6	410,473	833.3	490,477	90.4	74.4	16	1.6	8	12.69	60.86
S206	3,318.4	342,388	891.8	1,027,794	92.8	75.8	17	0.2	7	14.52	61.86
S227	3,629.8	404,099	865.7	948,877	93.5	79.5	14	0.7	5.8	14.19	60.41
S387	3,617.7	324,828	883.8	999,124	93.5	80	13.5	0.7	5.8	13.65	61.51
ZM3	2,652.8	649,952	805.5	1,389,025	94.6	85.1	9.5	0.7	4.7	14.74	59.91

generated an 826.16-Mb contig-level monoploid assembly with an N50 (the minimum contig length needed to cover 50% of the genome) of 7.25 Mb, the contigs of which were subsequently corrected and anchored in a chromosome-scale genome using the ALLHiC program. The final assembly comprised 826.07-Mb contigs (99.93%), anchored to eight monoploid pseudo-chromosomes. The assembly quality was evaluated using synteny analysis, a Hi-C contact matrix, benchmarking universal single-copy orthologs (BUSCO)[18] and the long terminal repeat assembly index (LAI)[19]. The Hi-C heatmap indicated that the entire genome can be divided into eight blocks, corresponding to eight chromosomes (Supplementary Fig. 2a). Sequence alignment and synteny analysis revealed strong collinearity among the genomes of ZM4 (Msa), *M. sativa* spp. *caerulea* (Mca) and *M. truncatula* genotype Jemalong A17 (Mt) (Supplementary Fig. 2b,c). Approximately 97.4% of complete BUSCOs (1,572 of 1,614) were detected in the monoploid genome. The LAI of the assembled monoploid genome was about 20.44, indicating that the ZM4 monoploid genome assembly reached the gold standard of Ou's classification system[19]. A total of 49,553 protein-coding genes were predicted from the ZM4 monoploid genome.

Variation detection and population structure analysis

To gain a deeper understanding of the genetic diversity and population structure of alfalfa, this study resequenced 176 core germplasm resources. We utilized admixture software to explore the global genetic structure of alfalfa, evaluating various models with the number of subpopulations (*K*) ranging from 2 to 9 (Supplementary Fig. 3). Cross-validation error rates were used to classify the 176 samples into three distinct subgroups: Group1, Group2 and Group3 (Extended Data Fig. 1a). Each subgroup consisted of representative accessions. More specifically, Group1 primarily included accessions from the United States, Group2 was predominantly composed of Chinese accessions and Group3 mainly consisted of accessions from Turkey (Extended Data Fig. 1b). The findings of a principal component analysis aligned with the results of the population structure assessment (Extended Data Fig. 1c), with the top three principal components explaining variances of 3.64%, 3.37% and 2.76%.

In the genetic analysis of different alfalfa subpopulations, the fixation index (F_{ST}) was used as a metric to measure genetic differentiation. Calculated F_{ST} values ranged from 0.011 to 0.018 among the three subpopulations (Extended Data Fig. 1d), indicating a low level of genetic differentiation between the subpopulations. This phenomenon may be influenced by relatively high gene flow, which could play a role in maintaining genetic similarity among the subpopulations. This research deepens our understanding of alfalfa's genetic diversity and provides valuable insights for its future genetic enhancement and the efficient utilization of alfalfa germplasm resources.

Assembly and annotation of 24 alfalfa accessions

To thoroughly explore alfalfa's genetic diversity and address potential gaps in the resequencing data, we selected 20 accessions from a set of 176 core germplasm resources. These samples were chosen based on phylogenetic relationships, geographical distribution and subgroup

https://doi.org/10.1038/s41588-025-02164-8

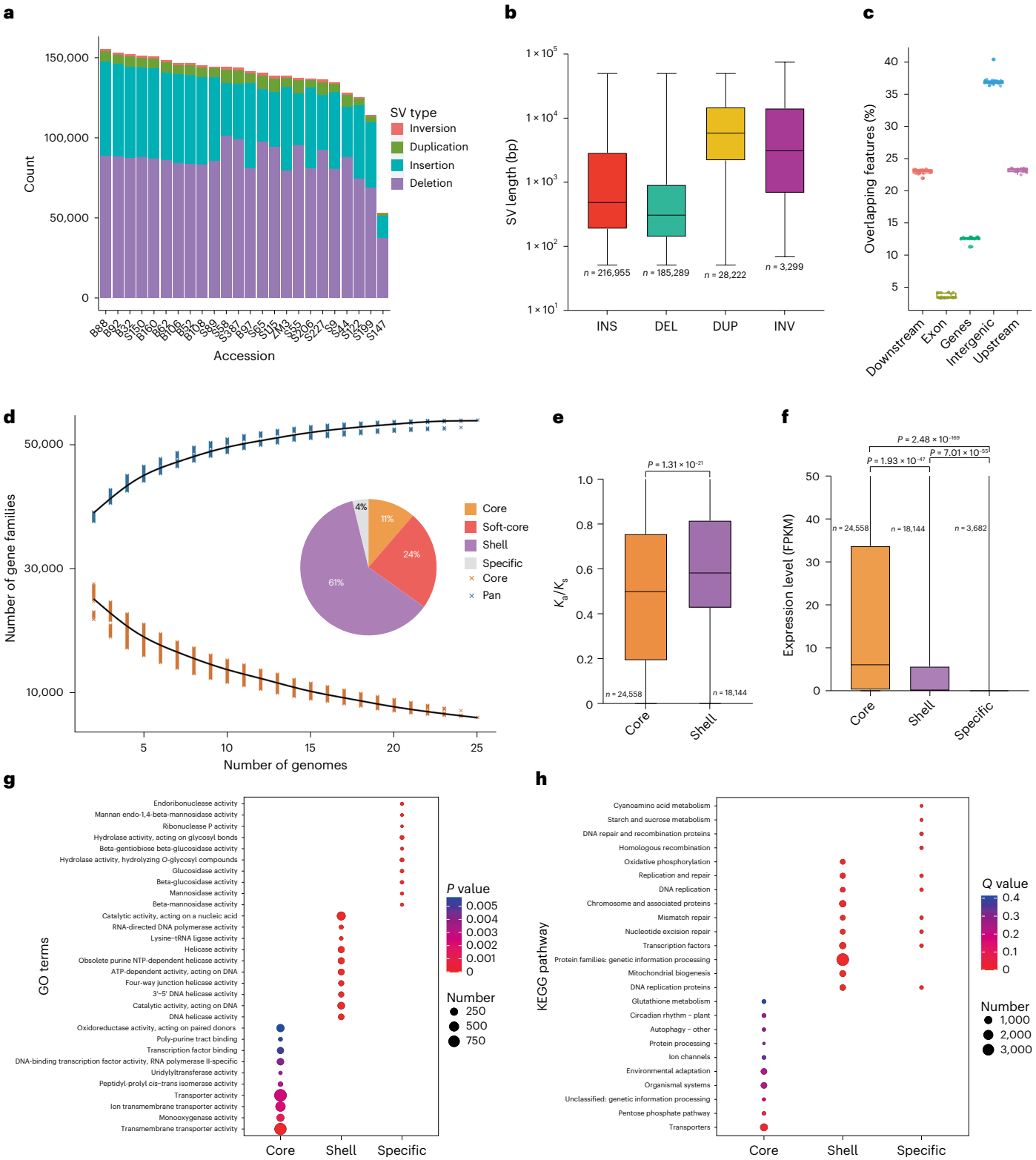


Fig. 3 | Detection of SVs and construction of the pan-genome based on the 24 de novo assembled alfalfa genomes. a, Count of different types of SVs across multiple samples. **b**, Length distribution of different SV types: insertion (INS), deletion (DEL), duplication (DUP) and inversion (INV). **c**, Percentage of overlapping features in the pan-genome analysis. The horizontal axis denotes different genomic regions, including downstream, exon, genes, intergenic and upstream. The vertical axis indicates the percentage of overlap of features in the pan-genome for each region. **d**, Proportion of genes categorized by genomic features (core, soft-core, shell, specific and pan). **e**, Distribution of the K_a/K_s ratio for core and shell genes. **f**, Expression level of core, shell and specific genes measured in fragments per kilobase of transcript per million mapped reads. **g**, Gene Ontology term enrichment for core, shell and specific genes with dot size representing the number of genes and color indicating the *P* value. **h**, KEGG, with dot size representing the number of genes. In boxplots, the 25% and 75% quartiles are shown as lower and upper edges of boxes, respectively, and central lines denote the median. The whiskers extend to 1.5× the interquartile range. Statistical significance is determined using bilateral Wilcoxon rank-sum tests. *n* represents the sample size in each group.

distributions. In addition, four representative accessions, not previously included in this population, were selected to ensure this study captured the broad spectrum of alfalfa's genetic variation; these four materials possessed unique heritage traits as well as distinct geographical origins (Fig. 2a,b). The collection comprised 24 representative alfalfa accessions selected for this study, including 2 wild accessions, 2 accessions with unclear improvement status, 11 landrace accessions and 9 cultivated accessions, spanning 5 continents and 18 countries and geographical areas (Fig. 2b and Supplementary Table 1). These selected germplasms exhibited not only extensive genetic diversity, but also notable phenotypic variations. The variations encompassed critical traits such as cold adaptation (fall growth), growth and development (leaf width), stress response (relative yield under salt stress) and nutritional quality (acid detergent fiber and in vitro for 24 h true dry matter degradability) (Fig. 2c,d). These comprehensive phenotypic observations can help to elucidate the underlying genetic basis and provide valuable information for future domestication and germplasm improvement research.

In this comprehensive study, an average of 37.2 Gb HiFi reads were generated for each of the 24 alfalfa accessions, effectively covering the haploid alfalfa genome size of approximately 800 Mb (Supplementary Table 2). These reads were assembled into initial contigs followed by the removal of redundant sequences. These assemblies represent monoploid genomes with sizes ranging from 801.7 Mb (B52) to 891.8 Mb (S206), with an average N50 contig size of 1.05 Mb (Table 1 and Supplementary Table 3). Chromosome lengths range from 799.2 to 881.53 Mb (Supplementary Table 4). Analysis of repetitive sequence proportions across 24 alfalfa accessions revealed a range from 55.11% (B108) to 63.20% (S58), with an average of approximately 59.93% (Table 1). These assemblies were further validated using 1,375 BUSCOs[18], reflecting an average completeness of 93.1%. All the BUSCO completeness scores exceeded 90%, reflecting a high level of completeness and a comprehensive representation of core genes in the assemblies (Table 1 and Supplementary Fig. 4). Furthermore, the values for complete and duplicated BUSCOs (D, %) in alfalfa ranged from 9.5% to 17%, which may reflect the genomic characteristics of alfalfa as a homotetraploid and suggest the presence of gene duplications in its genome (Table 1).

We also evaluated the LAI[19] of the assembled genomes, which ranged from 12.69 (S199) to 15.04 (B160) (Table 1). This range indicates that these assemblies have achieved a reference-level quality[19]. To further enhance the utility of our assemblies for downstream analyses, including genomic comparisons, gene synteny and population genetics studies, we anchored all contig-level assemblies to pseudo-chromosomal levels. This was accomplished using a reference-based method[20], yielding an average anchoring rate of 99.2% (Supplementary Table 3). These chromosomal-level assemblies not only exhibited similar chromosome sizes, but also demonstrated a high degree of synteny (Supplementary Fig. 5). However, a few inversions were noted, indicating the presence of SVs in specific regions of the alfalfa genomes (Extended Data Fig. 2).

Using ab initio and homology-based prediction methods, we identified a range of 47,587 to 53,837 protein-coding genes across different accessions, indicating variation in gene count among them (Supplementary Table 5). These gene counts closely align with those reported in previously published alfalfa genomes[2,4,5,21]. The average gene length in these genomes varied from 2,729.22 bp (the shortest, B62) to 2,906.88 bp (the longest, S150). For functional annotation, we utilized databases such as PFAM[22], eggNOG[23] and Kyoto Encyclopedia of Genes and Genomes (KEGG), successfully annotating approximately 93.0% of the genes on average.

SV detection and pan-genome construction

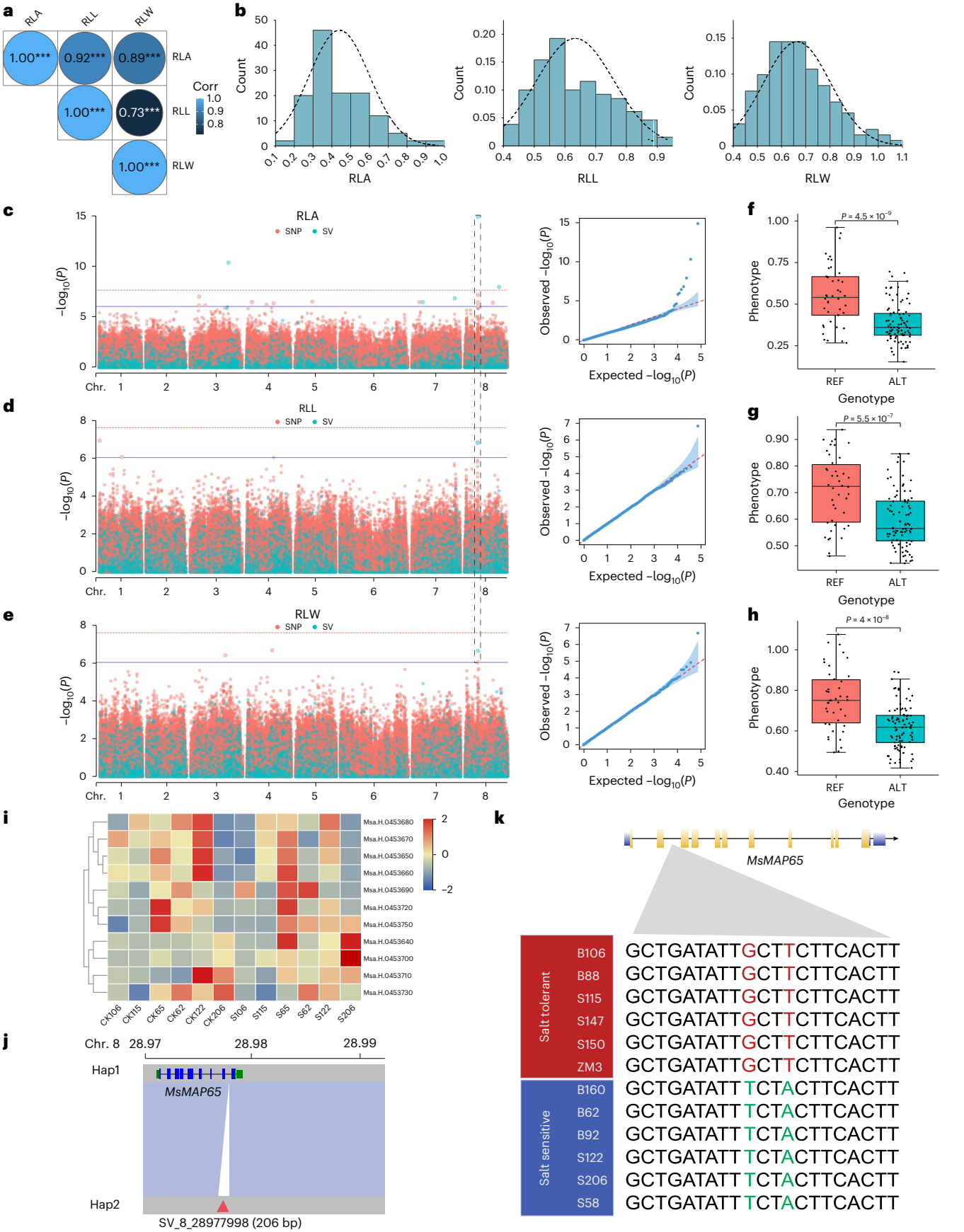
SVs contribute to the diversification of key biological characteristics across species. In exploring the genomic SV among different alfalfa accessions, this study pinpointed a total of 433,765 SVs, utilizing ZM4 alfalfa as the reference genome. These SVs fall into four categories: deletions, insertions, duplications and inversions (Fig. 3a, Extended Data Fig. 2 and Supplementary Table 6). On average, each variety exhibited 108,441 SVs. Notably, deletions and insertions were the predominant SV types, with 185,289 and 216,955 occurrences, respectively, while duplications and inversions were less common (Fig. 3a and Supplementary Table 6). Moreover, the distribution of these SV types was not uniform across chromosomes. For example, insertions were more frequent on chromosomes 1, 3 and 4, whereas inversions were more common on chromosomes 4 and 5 (Supplementary Table 6). These variations in distribution across chromosomes could reflect differences in their respective functions. In our analysis, significant differences in length were observed among the four SV types, specifically, with duplications and inversions being longer than deletions and insertions. These findings suggest that larger duplications and inversions may have a more substantial impact on genomic structure and function (Fig. 3a,b and Supplementary Table 6). Our analysis revealed a distinct pattern of overlap between SVs and various genomic features. There was a notably lower percentage of SVs overlapping with exon regions, suggesting that coding sequences may be more conserved under the influence of SV. Conversely, intergenic and upstream regions showed a higher percentage of overlap with SVs, which may indicate their association with genomic diversity (Fig. 3c).

To create a comprehensive gene catalog for alfalfa, we constructed a gene-based pan-genome by clustering 1,267,755 predicted gene models derived from 24 accessions and the reference genome ZM4. We used the Markov cluster algorithm, which identified 54,002 nonredundant gene families. The size of the pan-genome expanded with each additional individual included in the study, nearing a plateau at a sample size of 22 (Fig. 3d). This plateau suggests that our sampling effectively captures the full breadth of the alfalfa gene repertoire, thereby reflecting the genetic diversity of this species well. These gene families were classified into four categories: core (genes present in all 25 accessions (6,116, 11%)), soft-core (genes found in 22–24 accessions (12,697, 24%)), shell (present in 2–21 individuals (33,147, 61%)) and accession-specific clusters (2,042, 4%) (Fig. 3d). Core and soft-core genes collectively accounted for approximately 35% of the total gene families per accession (Fig. 3d), highlighting the potential for extensive genetic variation in the alfalfa population.

Gene Ontology and KEGG pathway analyses revealed that core genes were predominantly associated with vital biological pathways,

Fig. 4 | Functional impact of SV in alfalfa leaf morphology under salt stress. **a**, Phenotypic correlations among relative leaf area (RLA), relative leaf length (RLL) and relative leaf width (RLW) under salt stress. **b**, Distributions of RLA, RLL and RLW phenotypes. **c**–**e**, Manhattan plots for SNP-GWAS and SV-GWAS for three phenotypic traits: RLA (**c**), RLL (**d**) and RLW (**e**). The horizontal lines represent the genome-wide significance threshold after Bonferroni correction ($\alpha = \frac{0.05}{n}$, where 'n' is the total number of independent SNPs and effective SVs). The corresponding QQ plots for the SV-GWAS of each trait are shown on the right. **f**–**h**, Boxplots comparing the RLL (**f**), RLW (**g**) and RLA (**h**) phenotypes between the reference (REF) and alternative (ALT) groups under salt stress. The sample sizes for the REF and ALT groups are 59 and 117, respectively. **i**, Gene expression levels near the SV site across different samples. **j**, A 206 kb SV (SV_8_28977998), highlighted by a red triangle, located in the intronic region of the *MsMAP65* gene. **k**, Sequence alignment of *MsMAP65* allelic variants across different alfalfa cultivars, with salt-tolerant samples displaying the G genotype, while the T genotype is associated with salt sensitivity. In boxplots, the 25% and 75% quartiles are shown as lower and upper edges of boxes, respectively, and central lines denote the median. The whiskers extend to 1.5× the interquartile range. *P* values were calculated using a two-tailed Student's *t*-test. Corr, correlations.

https://doi.org/10.1038/s41588-025-02164-8



such as transmembrane transporters, ion channels and protein processing. By contrast, shell genes were more involved in biological processes related to genetic information processing, transcription factors and nucleotide repair ($P < 0.05$, Fisher's exact test) (Fig. 3g,h). Notably, core genes exhibited a significantly lower ratio of nonsynonymous to synonymous substitutions (K_a/K_s) compared with shell genes ($P < 1.31 \times 10^{-21}$, Wilcoxon rank-sum test), suggesting that core genes may be under stronger purifying selection and are more likely to be functionally conserved (Fig. 3e). Expression analysis, measured by fragments per kilobase of transcript per million mapped reads, showed that core genes had the highest mean expression levels. Shell genes exhibited moderate expression, while specific genes showed the lowest expression levels (Fig. 3f). This pattern suggests that core genes may play key roles in essential life processes, whereas specific genes could be associated with specialized or adaptive functions.

Graph-based genomic analysis of SV salt and quality traits

To detect SVs associated with agronomic traits, we mapped short-read sequences to the graph-based pan-genome, genotyping 54,649 valid SVs across 176 alfalfa germplasm resources for further analysis. First, we examined relative leaf length, width and area under salt stress, uncovering a strong correlation among these traits (correlation coefficients between 0.73 and 0.92) (Fig. 4a). These traits exhibited a normal distribution, laying a solid statistical foundation for genetic analysis (Fig. 4b). GWAS pinpointed an SV consistently associated with leaf traits under salt stress (Fig. 4c–e), suggesting that these SVs may encompass genes crucial for plant growth and morphological adaptation to salt stress. Haplotype analysis underscored the genetic variations linked to these adaptive responses (Fig. 4f–h). To further investigate the functional impact of these variations, we conducted a salinity tolerance evaluation and categorized accessions based on their response to salt stress (Supplementary Fig. 6). In the vicinity of this important SV, we identified 11 genes, and transcriptome analysis between three salt-tolerant and three salt-sensitive accessions revealed differential expression patterns for these genes under salt stress (Fig. 4i). Notably, one gene situated in the intronic region of this notable SV encodes the microtubule-associated protein gene *MsMAP65* (*Msa.H.0453710*) (Fig. 4j), which plays a crucial role in both development and the stress response[24,25]. A nonsynonymous mutation in the *MsMAP65* was correlated with salt tolerance, identifying it as a potential target for enhancing stress resilience in alfalfa (Fig. 4k).

In the fields of modern plant genetics and agricultural science, a profound understanding of primary phenotypic traits, especially those related to quality, is crucial for crop improvement[26,27]. Alfalfa, being an important forage crop, is no exception. In our comprehensive study of genetic traits associated with alfalfa quality, we identified a notable 233 bp SV on chromosome 1 (Extended Data Fig. 3). Specifically, this SV is significantly associated with monosaccharide content ($P < 2.44 \times 10^{-7}$) and closely relates to in vitro digestibility within 24 h ($P < 7.50 \times 10^{-7}$) and 30 h ($P < 4.42 \times 10^{-21}$). This suggests that the variation in SV may play a key role in regulating the digestibility and nutritional value of alfalfa. Through detailed haplotype analysis, we further revealed a close link between the genetic association of these traits and specific haplotype differences on chromosome 1 (Extended Data Fig. 3). This not only emphasizes the role of SV in shaping alfalfa quality, but also suggests the importance of genetic variation beyond SNPs.

Building on this understanding of genetic influences on alfalfa quality, we focused on another key phenotypic trait—the stem–leaf ratio (SLR). Among various phenotypic traits, the SLR, an indicator of the relative proportion of stem and leaf in a plant, plays a key role in assessing plant structural morphology and ecological adaptability. Previous research has shown a significant genetic correlation between SLR and biomass yield, suggesting that SLR can be an important indicator of yield estimation and quality[28,29]. In this study, we used SV-GWAS methodology to successfully identify SVs closely associated with the SLR phenotype (Fig. 5a). Haplotype analysis underscored the genetic variations linked to these developmental traits (Fig. 5b). A particularly notable finding during this process was the identification of eight genes in a 50-kb region upstream and downstream of this SV. These genes provide critical clues for further functional studies. Specifically, our analysis revealed that this SV is located downstream of the key gene, GA3-oxidase (*MsGA3ox1*, *Msa.H.0115520*), which plays a regulatory role in the gibberellin synthesis pathway and is crucial for plant growth and adaptability[30,31].

To evaluate the role of the candidate gene *MsGA3ox1* in alfalfa, we generated four independent overexpression lines (*MsGA3ox1-OE*) (Fig. 5c and Extended Data Fig. 4). Compared with wild-type plants, *MsGA3ox1* expression levels were significantly elevated in these transgenic lines (Fig. 5d and Extended Data Fig. 4b). Notably, *MsGA3ox1-OE* plants remained unchanged in height but showed slightly increased branching, along with enhanced leaf area, length and width, leading to sustained biomass and a marked reduction of 22–31% in the SLR (Fig. 5c–m). To validate the reliability of these findings, we conducted supplementary experiments using a different set of overexpression lines, OE7 and OE12, along with WT and OE3 grown in large outdoor pots for 2 months. These additional test results were essentially consistent with our initial observations in OE1 and OE3 grown in a greenhouse, further confirming the pronounced phenotypic changes caused by overexpression of *MsGA3ox1* (Extended Data Fig. 4). In addition, the number of trifoliolate leaves of *MsGA3ox1-OE* was significantly higher than that of WT (Extended Data Fig. 4j). The results showed that the decrease in SLR following overexpression of *MsGA3ox1* in alfalfa was attributed to the increase in trifoliolate leaves number and leaf size. Given that the SLR is closely related to forage quality, we also observed improvements in forage quality (Extended Data Fig. 5). These findings highlight *MsGA3ox1* as a key gene for reducing the SLR and enhancing forage quality, thereby playing a crucial role in alfalfa breeding programs.

Graph pan-genome empowers alfalfa breeding

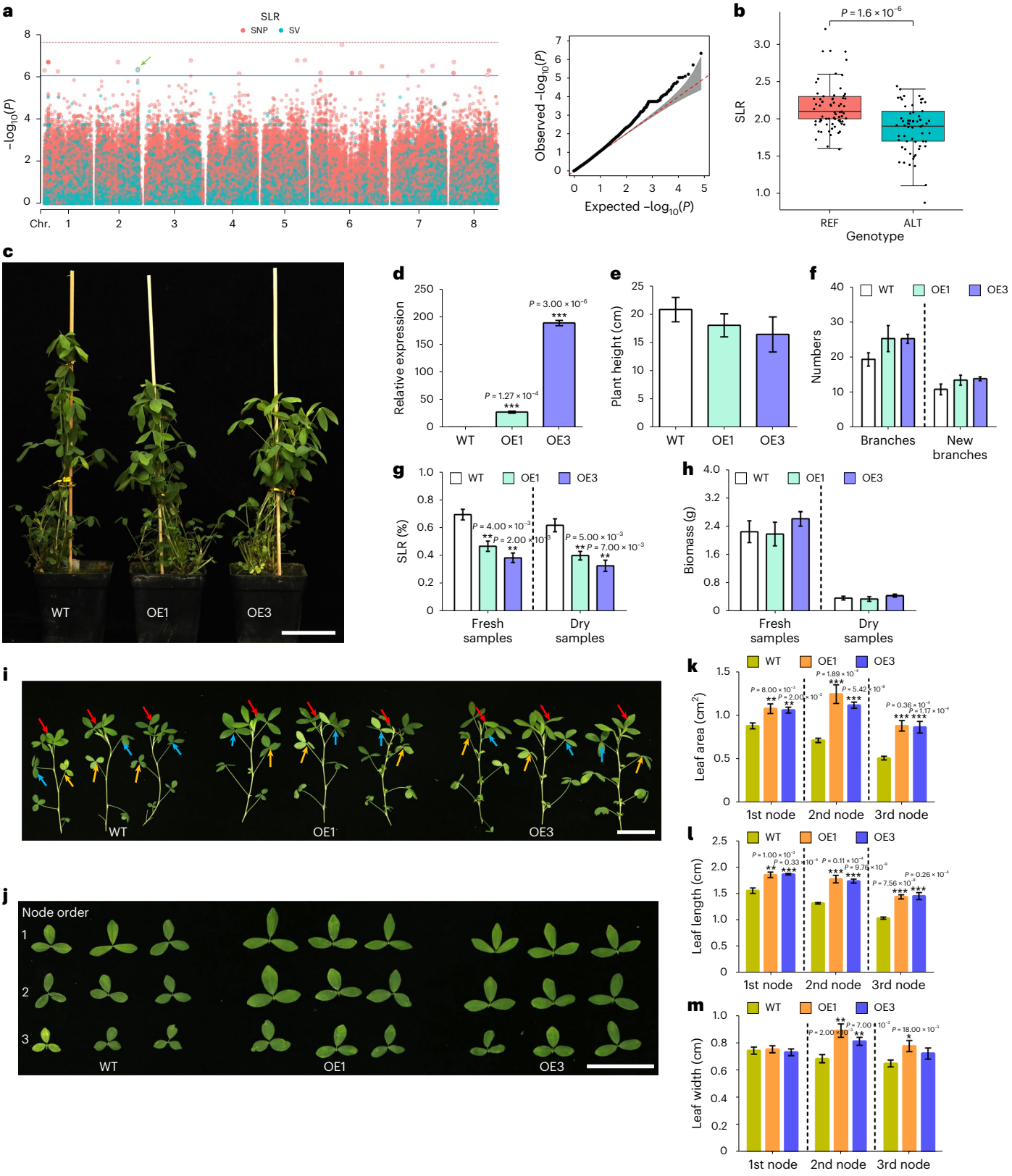
This study has successfully developed a graph-based pan-genome that integrates both reference and alternative alleles, while maintaining the coordinate framework of the linear reference genome. This comprehensive dataset enhances the mapping of short reads to SV regions, thus streamlining SV genotyping[32,33]. Leveraging second-generation sequencing data, we conducted comprehensive genotyping of 176 samples, which can serve as the training dataset for genomic selection (GS).

Fig. 5 | Functional validation of a key gene identified by pan-genomic analysis of SLR phenotype using SNP-GWAS and SV-GWAS. a, Manhattan plots for the SLR from SNP-GWAS and SV-GWAS, with horizontal lines indicating the genome-wide significance threshold Bonferroni correction ($\alpha = \frac{0.05}{n}$, where n is the total number of independent SNPs and effective SVs). Green arrows represent significant loci. The QQ plots for SV-GWAS are shown on the right. **b**, Distribution of SLR phenotypes between the REF and ALT genotype groups. The sample sizes for the REF and ALT groups are 102 and 74, respectively. **c**, Alfalfa plants grown in a greenhouse for one month after trimming. The control group (WT) is the Zhongmu No.1 variety of *M. sativa* L. **d**, Relative expression levels of *MsGA3ox1* in overexpression lines. **e**,**f**, Comparative analysis of plant height (**e**) and the number of branches (**f**), including new branches. **g**,**h**, SLR (**g**) and biomass (**h**) measurements between WT and transgenic alfalfa lines. **i**, Branches from WT, OE1 and OE3 lines. The red, blue and yellow arrows indicate leaves at the first, second and third stem nodes, respectively. **j**, Leaves from WT and overexpression lines (OE1 and OE3) corresponding to the first (1), second (2) and third (3) nodes in **i**. **k**–**m**, Comparative assessments of leaf area (**k**), leaf length (**l**) and leaf width (**m**) between WT and *MsGA3ox1* overexpression lines. Asterisks denote statistical significance: $^*P < 0.05$, $^{**}P < 0.01$ and $^{***}P < 0.001$. In boxplots, the 25% and 75% quartiles are shown as lower and upper edges of boxes, respectively, and central lines denote the median. The whiskers extend to 1.5× the interquartile range. Statistical significance is determined using bilateral Wilcoxon rank-sum tests. Data are presented as mean ± s.e.m., with three (**d**), four (**e**–**h**) or nine (**k**–**m**) independent experiments. Scale bars, 5 cm.

https://doi.org/10.1038/s41588-025-02164-8

Over five years, we collected 54 phenotypic datasets from Langfang (Hebei Province) and Yinchuan (Ningxia Hui Autonomous Region) of China, evaluating traits related to cold and salt tolerance, yield and hay quality. Our analysis revealed that most phenotypes were significantly influenced by the conditions of their outdoor growing environments (Fig. 6a, Supplementary Fig. 7 and Supplementary Table 7).

In an effort to enhance breeding capabilities across diverse environmental settings and more effectively harness genetic resources, we conducted comprehensive studies involving GWAS and GS across 54 distinct phenotypes. Our findings demonstrate a notable improvement in the efficacy of GWAS when incorporating SVs compared with traditional SNP-based approaches, particularly among certain traits.



https://doi.org/10.1038/s41588-025-02164-8

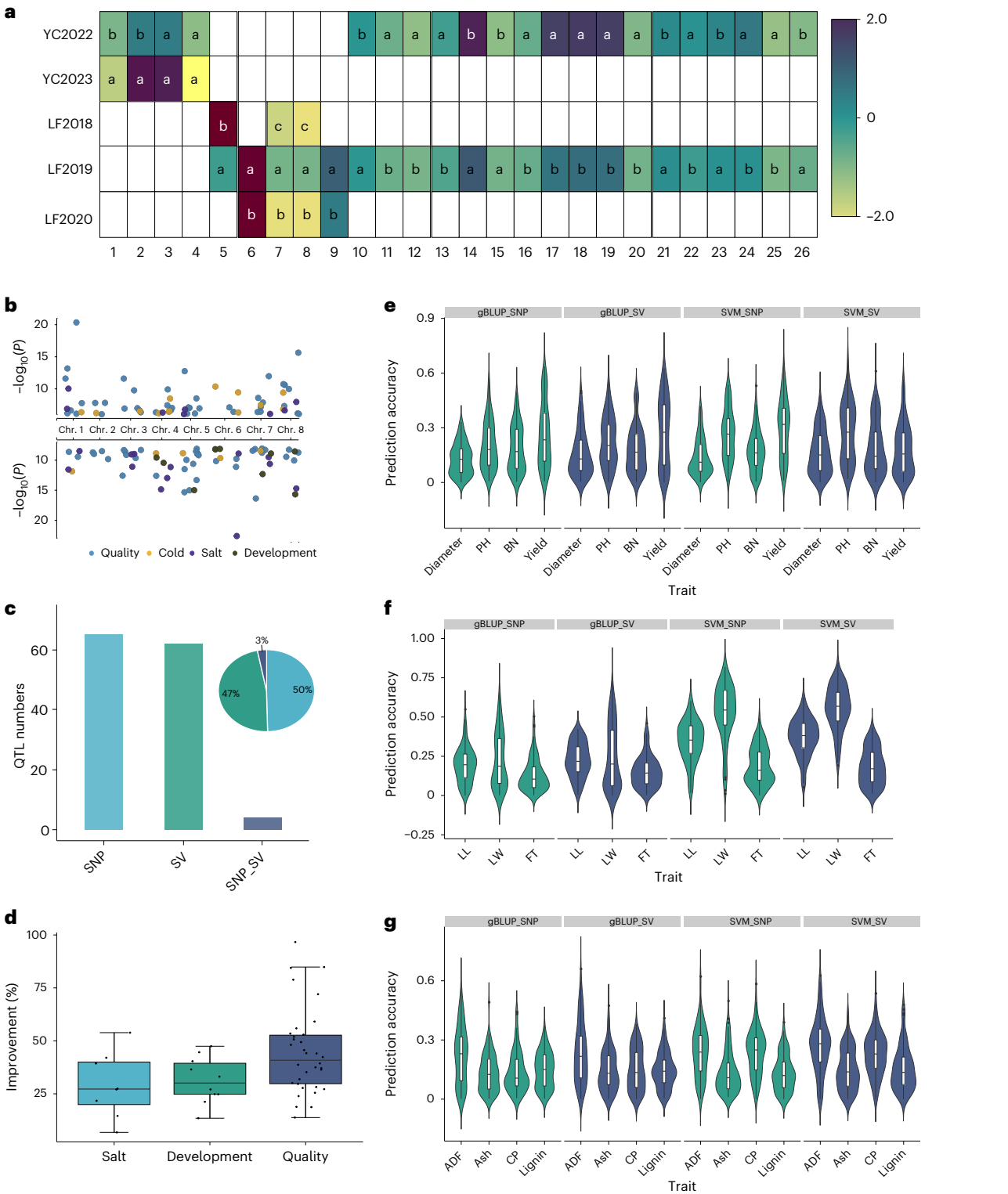


Fig. 6 | GWAS and genomic prediction accuracies using SV and SNP markers across 54 phenotypic traits. a, Phenotypic variation among different growth conditions. Different letters in the heatmap indicate significant differences ($P < 0.05$) as determined by pairwise two-sided Student's *t*-tests. Heatmap color represents the scaled phenotype values. **b**, Manhattan plots for significant SV-GWAS (upper) and SNP-GWAS (lower) across the 54 phenotypic traits, the bottom line represents the threshold for significance. **c**, Number of QTL detected by different types of genetic markers. SNP denotes QTLs identified solely by SNPs; SV denotes QTLs identified solely by SVs; SNP_SV denotes QTLs detected by both SNPs and SVs. **d**, Improvement percentage of salt ($n = 8$), development ($n = 10$) and alfalfa quality-related traits ($n = 34$) using base substitution of the top 20 highest effective variants. **e**–**g**, Comparison of the predictive accuracy of SNPs and SVs for various phenotypes using genomic BLUP and SVM models, including phenotypes under salt stress conditions (**e**), developmental-related phenotypes (**f**) (leaf length (LL), leaf width (LW) flowering time (FT)) and alfalfa quality-related phenotypes (**g**) (acid detergent fiber (ADF), ash, crude protein (CP) and lignin content). Each phenotype sample size was 100. Each violin plot includes a box plot representing data distribution: the center line indicates the median, the box edges represent the interquartile range (25th and 75th percentiles), and the whiskers extend to the minimum and maximum values within 1.5× the interquartile range. Statistical significance is determined using bilateral Wilcoxon rank-sum tests.

Specifically, 47% (59 of the detected signals or quantitative trait loci (QTL)) of the signals were uniquely identified through the application of SV-GWAS (Fig. 6b,c and Supplementary Table 8).

In this study, we utilized the results of SNP and SV for GS to predict phenotypes related to salt stress tolerance, developmental traits and alfalfa quality. Statistical analysis showed that, in predicting phenotypes under salt stress (Fig. 6e), developmental (Fig. 6f) and alfalfa quality (Fig. 6g), the accuracy of SV predictions was superior to that of SNP-based predictions in most cases. This result further emphasizes the importance of considering SVs in predicting complex traits. Compared with traditional SNP-based methods, the use of SVs demonstrated higher predictive accuracy, offering a new perspective on understanding and leveraging genetic diversity. To harness alfalfa's breeding potential, we calculated genomic estimated breeding values (GEBVs) for 52 selected phenotypic traits from an initial set of 54. Predictive accuracies ranged from 6.79% to 53.87% for salt stress traits, from 13.48% to 47.46% for developmental traits and from 13.80% to 96.63% for quality traits. These results suggested that GEBVs could substantially improve selection for salt stress, development and quality traits by up to 29.10%, 31.35% and 44.54%, respectively (Fig. 6d and Supplementary Table 9). Our findings highlight the transformative impact of pan-genome studies on crop breeding efficiency and open new avenues for agricultural advancement and genetic research.

Discussion

Alfalfa is one of the most palatable forages, offering high energy and protein content essential for dairy cows and other livestock. Despite challenges associated with genetic diversity and genome complexity, high-quality reference genomes and annotations are crucial for uncovering the genetic traits of this crop. In this study, we constructed a pan-genome of alfalfa, using 24 different accessions, aimed at exploring its genetic diversity and breeding potential. Given alfalfa's complexity as a tetraploid plant, we adopted a haploid-level genome assembly method, overcoming challenges posed by homologous chromosome sequence recombination and mismatches common in traditional methods. This approach, which utilizes bioinformatics tools to deconstruct the tetraploid genome into haploid-level sequences, simplified the genomic structure and minimized interference from repetitive sequences, laying a solid foundation for the accurate identification of genes associated with important agronomic traits.

Further, this study extensively explored the key role of SVs in genetic diversity and genome functionality. We discovered that the main types of SVs in alfalfa, deletions and insertions, resemble those observed in other plant species such as rice[34], tomato[35] and *Setaria italica*[36], suggesting these genomic variations might play a substantial and conservative role in plant evolution. Comparative analysis of the pan-genome with other species revealed that only 11% of alfalfa's core genome is shared, substantially lower than in oranges[37], *Setaria italica*[36] and soybeans[38], reflecting its unique genetic structure and population diversity. In addition, we found that the ratio of nonsynonymous to synonymous substitutions (K_a/K_s) in core genes was significantly lower than in shell genes, consistent with findings in other species[12], emphasizing the importance of maintaining genetic stability in the conserved regions of the genome.

In this study, we also successfully constructed a graph-based alfalfa pan-genome, overcoming the limitations of traditional short-read sequencing in capturing large SVs. This method revealed additional genetic variations in alfalfa under salt stress and quality, and provided new perspectives for understanding its adaptability in various environments and quality. Of particular interest, we identified two genes, *MsMAP65* and *MsGA3ox1*, playing crucial roles in alfalfa adaptability and quality. *MsMAP65* is associated with salt tolerance in *Arabidopsis* and cucumber[24,25], highlighting its importance in the environmental stress response. Similarly, *MsGA3ox1*, which influences the gibberellin pathway, underscores the significance of SV in the regulation of leaf development and plant growth. Previous studies have shown that mutants of *GA3ox* in *Arabidopsis* and *M. truncatula* and *M. sativa* result in significantly dwarfed plants with smaller leaves[39–41]. However, overexpression of the *Arabidopsis* gene in aspen trees did not lead to notable differences in plant height[42], whereas overexpression of *PsGA3ox1* in peas resulted in larger stipuless[30]. This is consistent with the result that overexpression of *MsGA3ox1* in alfalfa did not significantly alter plant height, but did significantly increase the number and size of trifoliolate leaves, leading to a reduction in SLR and an increase in quality. In addition, previous studies have reported that the *MsGA3ox1* mutant can also increase the leaf–stem ratio; however, this occurs at the expense of plant height and internode length, resulting in a plant dwarf[43]. These findings deepen our understanding of the role of the *MsGA3ox1* gene in plant growth regulation and quality improvement, paving new directions for future alfalfa breeding.

This study highlights the important impact of the growth environment on the expression of plant traits, revealing the complex interplay between genetic factors and external conditions. By comparing GWAS based on SNPs with those based on SVs, we identified key phenotypic loci with minimal overlap, indicating that SVs capture a rich set of genetic information inaccessible through SNP markers. This finding provides a new perspective for understanding genetic diversity and paves the way for future breeding strategies. Moreover, our research discovered that using SV markers alone results in higher predictive accuracy for most traits compared with using only SNP markers, highlighting the important role of SV markers in predicting complex genetic traits. However, the overall predictive accuracy observed in our study was relatively low. This could potentially be attributed to factors such as the limited sample size. Upon reviewing similar studies, it is apparent that this challenge is not uncommon, suggesting that increasing the sample size in future studies might be one of several approaches to potentially improve predictive accuracy[44]. Despite the limitations in predictive accuracy, our research provides valuable insights into the genetic mechanisms underlying complex traits. It illuminates new pathways and poses fresh challenges for future genetic analysis and crop breeding, emphasizing the need for continued exploration in this critical field.

In conclusion, this research not only broadens our understanding of the genetic diversity of alfalfa, but also provides new strategies for breeding, particularly to enhance crop adaptability and quality. The application of pan-genome analysis highlights its capacity to reveal genetic variations at the whole-genome level, which is crucial for adapting to global changes and improving crop traits. Future research should focus on exploring the role of SVs in biological adaptability and stress response, integrating this genetic information into breeding programs to enhance crop yield, quality and environmental adaptability.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-025-02164-8.

References

1. Annicchiarico, P., Barrett, B., Brummer, E. C., Julier, B. & Marshall, A. H. Achievements and challenges in improving temperate perennial forage legumes. *Crit. Rev. Plant Sci.* **34**, 327–380 (2015).
2. Shen, C. et al. The chromosome-level genome sequence of the autotetraploid alfalfa and resequencing of core germplasms provide genomic resources for alfalfa research. *Mol. Plant* **13**, 1250–1261 (2020).
3. Li, X. & Brummer, E. C. Applied genetics and genomics in alfalfa breeding. *Agronomy* **2**, 40–61 (2012).
4. Chen, H. et al. Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. *Nat. Commun.* **11**, 2494 (2020).

5. Long, R. et al. Genome assembly of alfalfa cultivar Zhongmu-4 and identification of SNPs associated with agronomic traits. *Genomics Proteomics Bioinformatics* **20**, 14–28 (2022).
6. Jayakodi, M., Schreiber, M., Stein, N. & Mascher, M. Building pan-genome infrastructures for crop plants and their use in association genetics. *DNA Res.* **28**, dsaa030 (2021).
7. Pang, A. W. et al. Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.* **11**, R52 (2010).
8. Zhang, Z. et al. Genome-wide mapping of structural variations reveals a copy number variant that determines reproductive morphology in cucumber. *Plant Cell* **27**, 1595–1604 (2015).
9. Zhou, Y. et al. The population genetics of structural variants in grapevine domestication. *Nat. Plants* **5**, 965–979 (2019).
10. Saxena, R. K., Edwards, D. & Varshney, R. K. Structural variations in plant genomes. *Brief. Funct. Genomics* **13**, 296–307 (2014).
11. Gabur, I., Chawla, H. S., Snowdon, R. J. & Parkin, I. A. Connecting genome structural variation with complex traits in crop plants. *Theor. Appl. Genet.* **132**, 733–750 (2019).
12. Chen, S. et al. Gene mining and genomics-assisted breeding empowered by the pangenome of tea plant *Camellia sinensis*. *Nat. Plants* **9**, 1986–1999 (2023).
13. Gaut, B. S., Seymour, D. K., Liu, Q. & Zhou, Y. Demography and its effects on genomic variation in crop domestication. *Nat. Plants* **4**, 512–520 (2018).
14. Wellenreuther, M., Mérot, C., Berdan, E. & Bernatchez, L. Going beyond SNPs: the role of structural genomic variants in adaptive evolution and species diversification. *Mol. Ecol.* **28**, 1203–1209 (2019).
15. Huang, K. & Rieseberg, L. H. Frequency, origins, and evolutionary role of chromosomal inversions in plants. *Front. Plant Sci.* **11**, 296 (2020).
16. Kirkpatrick, M. & Barton, N. Chromosome inversions, local adaptation and speciation. *Genetics* **173**, 419–434 (2006).
17. Zhang, X. et al. Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant *Camellia sinensis*. *Nat. Genet.* **53**, 1250–1259 (2021).
18. Simão, F. A., Waterhouse, R. M., Panagiotis, I., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
19. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
20. Alonge, M. et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 258 (2022).
21. Li, A. et al. A chromosome-scale genome assembly of a diploid alfalfa, the progenitor of autotetraploid alfalfa. *Hortic. Res.* **7**, 194 (2020).
22. Finn, R. D. et al. Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).
23. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
24. Zhou, S., Chen, Q., Li, X. & Li, Y. MAP65-1 is required for the depolymerization and reorganization of cortical microtubules in the response to salt stress in *Arabidopsis*. *Plant Sci.* **264**, 112–121 (2017).
25. Liang, M. et al. Comprehensive analyses of microtubule-associated protein MAP65 family genes in Cucurbitaceae and *CsaMAP65s* expression profiles in cucumber. *J. Appl. Genet.* **64**, 393–408 (2023).
26. Dwiningsih, Y. & Al-Kahtani, J. Genome-wide association study of complex traits in maize detects genomic regions and genes for increasing grain yield and grain quality. *Adv. Sustain. Sci. Eng. Technol.* **4**, 0220209 (2022).
27. Liu, R. et al. GWAS analysis and QTL identification of fiber quality traits and yield components in upland cotton using enriched high-density SNP markers. *Front. Plant Sci.* **13**, 1067 (2018).
28. Kephart, K. D., Buxton, D. & Hill, R. Jr Digestibility and cell-wall components of alfalfa following selection for divergent herbage lignin concentration. *Crop Sci.* **30**, 207–212 (1990).
29. Han, R.-H., Lu, X.-S., Gao, G.-J. & Yang, X.-J. Analysis of the principal components and the subordinate function of alfalfa drought resistance. *Acta Agrestia Sin.* **14**, 142 (2006).
30. Reinecke, D. M. et al. Gibberellin 3-oxidase gene expression patterns influence gibberellin biosynthesis, growth, and development in pea. *Plant Physiol.* **163**, 929–945 (2013).
31. Wu, H., Bai, B., Lu, X. & Li, H. A gibberellin-deficient maize mutant exhibits altered plant height, stem strength and drought tolerance. *Plant Cell Rep.* **42**, 1687–1699 (2023).
32. Ameur, A. Goodbye reference, hello genome graphs. *Nat. Biotechnol.* **37**, 866–868 (2019).
33. Garrison, E. et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
34. Qin, P. et al. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* **184**, 3542–3558.e16 (2021).
35. Alonge, M. et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**, 145–161.e23 (2020).
36. He, Q. et al. A graph-based genome and pan-genome variation of the model plant *Setaria*. *Nat. Genet.* **55**, 1232–1242 (2023).
37. Huang, Y. et al. Pangenome analysis provides insight into the evolution of the orange subfamily and a key gene for citric acid accumulation in citrus fruits. *Nat. Genet.* **55**, 1964–1975 (2023).
38. Liu, Y. et al. Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162–176.e13 (2020).
39. Hu, J. et al. Potential sites of bioactive gibberellin production during reproductive growth in *Arabidopsis*. *Plant Cell* **20**, 320–336 (2008).
40. Sun, H. et al. Gibberellins inhibit flavonoid biosynthesis and promote nitrogen metabolism in *Medicago truncatula*. *Int. J. Mol. Sci.* **22**, 9291 (2021).
41. Dalmadi, Á. et al. Dwarf plants of diploid *Medicago sativa* carry a mutation in the gibberellin 3-β-hydroxylase gene. *Plant Cell Rep.* **27**, 1271–1279 (2008).
42. Israelsson, M., Mellerowicz, E., Chono, M., Gullberg, J. & Moritz, T. Cloning and overproduction of gibberellin 3-oxidase in hybrid aspen trees. Effects on gibberellin homeostasis and development. *Plant Physiol.* **135**, 221–230 (2004).
43. Zheng, L. et al. From model to alfalfa: gene editing to obtain semidwarf and prostrate growth habits. *Crop J.* **10**, 932–941 (2022).
44. He, X. et al. Accuracy of genomic selection for alfalfa biomass yield in two full-sib populations. *Front. Plant Sci.* **13**, 1037272 (2022).

Methods

Sample selection and sequencing

We selected 176 germplasm resources based on ref. 45, which provided a comprehensive evaluation of these materials. These accessions originated from 45 countries, including China, the United States and Turkey, with geographical distribution spanning Europe, Asia, Africa and the Americas. All samples were sequenced using the BGI platform and included 14 wild materials, 73 landrace accessions, 72 cultivated accessions and 17 accessions with uncertain improvement status. Paired-end sequencing reads were aligned to the ZM4 reference genome using BWA-MEM[46], and approximately 29.6 million SNPs were detected via the SAMtools[47] VarScan pipeline. The data were filtered using vcftools (v.0.1.16)[48] based on the criteria of a missing rate ≤10%, a minimum average read depth >5 and a minor allele frequency of >0.05. To explore evolutionary relationships, high-quality SNP loci were extracted to generate a SNP matrix and a phylogenetic tree was constructed using fasttree[49] with parameters -nt --gtr. Population structure was analyzed with admixture[50] software, with testing various subgroup numbers (*K* values) and determining the optimal *K* via cross-validation.

From this dataset, 20 alfalfa accessions were selected based on phylogenetic relationships, geographical distribution and subgroup classifications to represent a broad genetic spectrum. In addition, 4 accessions not included in the initial 176 were incorporated for their unique heritage traits and distinct geographical origins, further enriching the representation of alfalfa's genetic diversity. These 24 accessions were sequenced on the PacBio Sequel II platform using circular consensus sequencing mode, generating 25.13–45.14 Gb of HiFi reads per sample (https://github.com/PacificBiosciences/ccs).

Reference genome assembly

A total of 2.74 Gb contigs were previously assembled for an allele-aware genome of ZM4 alfalfa[5]. In this study, the contigs were used to assemble the haplotype genome after removing duplication using the Khaper (Kmer-based HaplotypeCaller) algorithm (v.1.0)[17]. The obtained nonredundant contigs were then interrupted and corrected by Juicer (v.1.5)[51] and 3D-DNA (v.180419) with Hi-C data. The corrected contigs were finally anchored to monoploid pseudo-chromosomes by ALLHiC[52,53] (https://github.com/tangerzhang/ALLHiC) and GMAP (v.2013-10-28)[54]. The genomes of *M. sativa* spp. *caerulea*[21] and *M. truncatula* ecotype Jemalong A17 (ref. 55) were used for synteny analysis with ZM4 by MCScanX[56].

Genome assembly of 24 alfalfa accessions

For the other 24 accessions, we assembled the genomes of 24 HiFi-sequenced accessions using hifiasm[57] (https://github.com/chhylp123/hifiasm; v.0.19.8-r603), with the default settings. The initial output of hifiasm yielded a primary assembly (p_ctg) that represented a composite haplotype with retained redundant sequences. To tailor the assemblies for downstream analysis, we initiated a de-duplication process that refined the monoploid genome assemblies by removing redundancies and enhancing the contiguity of primary contigs. We excluded collapsed haplotigs from the primary assembly using the default parameters of Khaper (v.1.0) in the 24 assemblies[17]. Furthermore, we used purge_dups (v.1.01)[58] to further streamline the de-duplication process. Assembly completeness was evaluated using the viridiplantae_odb10 database from BUSCO (v.5.0)[18] with its default settings.

For genomes absent of Hi-C sequencing data, chromosome-scale assemblies were attained via a reference-guided strategy. This entailed aligning contigs against the monoploid ZM4 reference genome, followed by the ordering and orientation of these contigs. The entire scaffolding procedure was performed using RagTag (v.1.1.1)[20] with the default parameters.

Genome annotation

We used RepeatModeler (v.1.0.11) to perform a comprehensive search for repetitive elements in each genome assembly. To accurately identify intact long terminal repeat retrotransposons (LTR-RTs), we integrated LTR_FINDER (v.1.0.7)[59] with default settings and LTR_harvest (v.1.5.10)[60] with finely tuned parameters: 'gt suffixerator -db ref -indexname ref -tis -suf -lcp -des -ssp -sds --dna' and 'gt ltrharvest -index ref -similar 90 -vic 10 -seed 20 -seqids yes -minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1'. Subsequently, LTR_retriever (v.1.6)[61] was used to eliminate any spurious intact LTR-RTs and to curate a streamlined, nonredundant LTR-RT library. Finally, RepeatMasker (v.1.332)[62] was utilized to mask repetitive sequences, incorporating the nonredundant libraries generated by both RepeatModeler and LTR_retriever.

For each alfalfa genome assembly, protein-coding genes were predicted using a dual approach comprising ab initio- and homology-based methods. Soft-masked genome assemblies provided the foundation for downstream gene prediction tasks. We engaged the ProtHint pipeline (v.2.6.0) to map protein sequences onto the genomes, providing essential hints for BRAKER (v.2.1.6)[63] to infer gene models using AUGUSTUS (v.3.3.0)[64] and to refine the training model with GeneMark-ES/ET (v.3.67)[63], incorporating '-softmasking' as a parameter. We curated a homologous protein sequence database from Phytozome13, featuring species such as *M. truncatula*, *Cicer arietinum*, *Glycine max*, *Prunella vulgaris*, *Prunus persica* and *Arabidopsis thaliana*, to enhance our homology-based prediction. In compiling final gene annotations, we selected the longest transcript for each gene model as the canonical representation. The quality of gene models was validated against their sequence similarity to known homologs. Our pipeline, encapsulated in a bespoke PERL script, further refined gene predictions. Utilizing the PFAM[22] and plant-specific UniProt databases, we screened for sequences with high conservation. Gene models were retained if they: (1) exhibited homology to proteins in the UniProt database or conserved PFAM domains; and (2) did not overlap with transposable element (TE) sequences. Protein functions and domains were annotated using InterProScan5 (ref. 65) and EggNOG-mapper[23].

Identification of TEs

To inform a comprehensive identification of TEs in each alfalfa genome assembly, we used the extensive de novo TE annotator EDTA (v.2.0.0)[66] package with default parameters. EDTA is a versatile tool that performs automated whole-genome de novo TE annotation, integrating both structure and homology approaches to generate high-quality nonredundant TE annotations. Before executing EDTA, we identified intact LTR-RTs using LTRharvest[60] and LTR_FINDER[59]. The identified LTRs underwent quality control through LTR_retriever[61]. Subsequently, terminal-inverted repeats were annotated using TIR-Learner (v.2.5)[67], and Helitron elements were identified with HelitronScammer (v.1.1)[68]. The remaining TE components were identified using RepeatModeler (v.2.0)[69] and homology-based annotation with RepeatMasker (v.1.332)[62], incorporating the EDTA TE library. The combined results of these annotation steps constituted the final TE annotation, ensuring a comprehensive and accurate representation of TEs in the alfalfa genome assemblies.

Pan-genome analysis

Reciprocal BLASTP[70] analyses were performed on 1,267,755 protein sequences derived from 24 alfalfa accessions and the 'ZM4' reference genome. The resulting data were processed using OrthoFinder (v.2.5.2)[71], which used the Markov cluster algorithm to cluster protein-coding genes, yielding 54,002 nonredundant pan-gene clusters. These clusters were stratified into core (present in all 25 genomes), soft-core (23–24 genomes), shell (2–22 genomes) and specific (unique to one genome). We modeled the pan-genome size using a random algorithm, varying the number of genome combinations at each iteration. The nonsynonymous to synonymous substitution ratio (K_a/K_s) in core and shell gene families was calculated using KaKs_Calculator (v.19)[72] with the MYN[73] method. We used Fisher's exact tests for multiple

comparisons to discern statistically significant variances. For functional enrichment, Fisher's exact tests were executed in R software, with a false discovery rate (FDR) threshold of <0.05 demarcating significant enrichment across functional classes.

Identification of SVs

To detect SVs across 24 assembled alfalfa genomes, we deployed two strategies: read-mapping and whole-genome alignment-based approaches. For read-mapping, corrected PacBio CLR reads for each genome were aligned to the ZM4 reference genome using NGLMR (v.0.2.7)[74] with default parameters. Subsequent SV detection, including deletions, insertions, duplications and inversions, was performed using SVIM (v.2.0.0)[75], cuteSV (v.2.0.3)[76] and Sniffles (v.1.0.12)[74]. We applied stringent filtering criteria to the detected SVs: a minimum support of five reads, size ranging from 50 bp to 50 kb, inclusion of a 'pass' tag and corroboration by at least two SV callers. High-confidence SVs were then merged using SURVIVOR (v.1.0.7)[77] with the parameters '1000 2 0 0 0 50', permitting a 1,000-bp maximum distance for merging and requiring dual tool validation for each SV.

In the whole-genome alignment-based approaches, NUCMER[78] facilitated the alignment of the 24 genomes against the ZM4 reference. Merging of alignment files was conducted again with SURVIVOR. The final integration of SVs from both read-mapping and whole-genome alignment-based approaches was executed using SURVIVOR (v.1.0.7)[77], culminating in a consolidated dataset of SVs representative of all samples.

Construction of the graph genome and SV genotyping

This analysis of 24 newly assembled alfalfa genomes yielded a high-quality dataset of SVs, setting the stage for a graph-based pan-genome construction. To ensure accuracy, we refined the SV dataset using BCFTools (v.1.13)[79], which involved realigning SV regions to the reference genome to rectify discrepancies in size and location. Leveraging the Variant Graph (vg) toolkit (https://github.com/vgteam/vg, v.1.34.0), we indexed the reference genome alongside variant data using the 'vg autoindex' command. The raw data were then converted into an xg formatted Variant Graph via 'vg convert', and the pan-genome was assembled using 'vg construct'.

For genotyping, resequencing reads from 176 alfalfa samples were mapped to the graph genome with 'vg giraffe'. We conducted coverage analysis on these mappings using 'vg pack', which converted the results into a binary representation. Variant calling on these data was executed with 'vg call', from which we extracted genotype information for each sample. The 'tabix' tool was then utilized to index each set of genotype data. In the final integration phase, we amalgamated the genotype data from all samples into a comprehensive file suitable for population analysis. This integrative approach synthesizes genetic data across samples, providing a robust basis for subsequent studies on population structure and genetic diversity.

Genome-wide association studies

Our GWAS comprised an extensive analysis of resequencing data from 176 carefully curated alfalfa samples, selected to represent genetic and phenotypic variation across diverse growth conditions. We investigated 80 phenotypic traits, covering critical agronomic attributes from stress responses and developmental phases to crop quality, aiming to capture the full spectrum of alfalfa's genetic and phenotypic diversity. The ZM4 genome, known for its broad representation and validation, served as our reference for precise genetic variation mapping.

Through rigorous data processing and analysis, we identified 2,043,025 SNPs and 54,649 SVs, genetic markers potentially crucial for trait determination. We conducted GWAS using the sophisticated FarmCPU model from the rMVP package (https://github.com/xiaolei-lab/rMVP), which explores the genetic underpinnings of complex traits while accounting for population structure. We also extended its application to SV-GWAS for a more comprehensive analysis of SVs.

To ensure the accuracy of our results, we applied the Bonferroni correction separately for the count of independent SNPs and effective SVs, setting a stringent genome-wide significance threshold of $\frac{0.05}{n}$. Here, 'n' represents the total independent SNPs and effective SVs. This approach allows a robust measure of statistical significance. The final GWAS findings were illustrated using the CMplot package (v.4.5.1) in R (v.4.3.1), offering clear and concise visualizations that highlight the key discoveries of our research.

Salt tolerance evaluation

To assess the salinity tolerance of alfalfa, we evaluated membership function values in the framework of fuzzy set theory[80]. This approach extends the classical concept of the indicator function, quantifying the degree to which a specific evaluative aspect holds true. By utilizing this methodology, we gain a nuanced and dynamic understanding of alfalfa's adaptability to salinity, allowing an evaluation that spans a continuous spectrum rather than a binary 'yes' or 'no' judgment. This methodological approach offers a comprehensive and sophisticated means to assess salinity tolerance in alfalfa with greater precision. By utilizing membership functions, this approach quantifies the salt tolerance of different alfalfa accessions, where each membership function maps elements of any set X to a real number in the interval [0, 1]. The MFV (membership function value) enables a more flexible classification of alfalfa accessions into the salt-tolerant category for assessment. Following standard procedures, we calculated the membership function value for salt tolerance using a specific equation, thereby allowing a comprehensive evaluation of the salt tolerance of alfalfa accessions.

GS prediction

In our GS analysis, we used two computational models to predict phenotypic outcomes: (1) genomic BLUP a predictive model that relies on a kinship-based relationship matrix, projects phenotypic traits by assessing genetic similarities among individual specimens, following the conceptual framework introduced in ref. 81; and (2) Support Vector Machine (SVM), selected for its robustness in binary classification scenarios. SVM was implemented following the optimization algorithm outlined in ref. 82, which aims to maximize the class separation margin. To validate the predictive performance of these models, we partitioned the dataset into two subsets: 80% of the samples were used for the training set and the remaining 20% formed the validation set.

Breeding potential prediction

In our study, we measured traits related to development, salt stress and quality, and calculated the GEBV for each trait. For predicting breeding potential, we focused on the top 20 most influential GEBVs. These selected GEBV markers enable a more accurate prediction and provide a deeper understanding of the breeding potential for various traits. To comprehensively evaluate the overall performance of the top accessions, we used the following formula to determine the percentage increase for each phenotype: $\frac{(\mathrm{GEBV}_{\mathrm{maxhaplotype}}-\mathrm{GEBV}_{\mathrm{maxcultivated}})}{\mathrm{GEBV}_{\mathrm{maxcultivated}}}\times 100\%$. This method evaluates the overall performance of a group of top accessions, providing a more comprehensive assessment of breeding potential.

Functional characterization of *MsGA3ox1*

A full-length coding sequence of *MsGA3ox1* was obtained through polymerase chain reaction amplification (primers: *MsGA3ox1*-F and *MsGA3ox1*-R) (Supplementary Table 10) using total RNA reverse transcription complementary DNA of *M. sativa* L. cv. Zhongmu-1 as a template. For overexpression vector construction, *MsGA3ox1* plasmid with the correct sequence was amplified using primers *MsGA3ox1*-1300-F and *MsGA3ox1*-1300-R (Supplementary Table 10), then homologously recombined with the expression vector Super1300 containing the preserved SmaI restriction site in our laboratory. Next, the confirmed

MsGA3ox1 overexpression plasmid was transferred into *Agrobacterium* EHA105 competent cells, and the subsequent *Agrobacterium*-mediated genetic transformation of alfalfa was carried out according to Fu's protocol[83]. Four independent overexpressed-*MsGA3ox1* (OE) alfalfa lines were obtained after hygromycin screening, DNA (primers: 1300-F and *MsGA3ox1*-1300-R) (Supplementary Table 10) and RNA-level identification (primers: *MsGA3ox1*-RT-F and *MsGA3ox1*-RT-R) (Supplementary Table 10). All alfalfa plants used for subsequent phenotypic characterization and verification in this experiment were obtained through asexual reproduction of OE lines and wild-type, respectively. They were soil cultured in a greenhouse under a 16-hour light (24 °C)/8-hour dark (20 °C) cycle, 65 mmol m^{-2} s^{-1} light intensity and 60–70% ambient humidity.

The plants, generated through asexual reproduction over the same period and regrown for 1 month after multiple mowing to a 3 cm height, were used to assess growth-related indicators. The distance from the top of the longest branch to the bottom of the branch node was measured as the plant height. Branches with new buds were counted as new branches. The aboveground leaf and stem tissues were harvested for fresh leaf weight (FLW) and fresh stem weight (FSW) measurement, respectively. Subsequently, fresh samples were oven-dried at 65 °C for 48 h for dry leaf weight (DLW) and dry stem weight (DSW) measurement, respectively. For the measurement of size-related indicators, all leaves were sampled from the same part of branches at the same height in both wild-type and transgenic lines for comparison. Leaf area, leaf length and leaf width were measured using ImageJ software[84]. For calculation,

$$\text{SLR (fresh samples)} = \frac{\text{FSW}}{\text{FLW}}$$

$$\text{SLR (dry samples)} = \frac{\text{DSW}}{\text{DLW}}$$

$$\text{Biomass (fresh samples)} = \text{FLW} + \text{FSW}$$

$$\text{Biomass (dry samples)} = \text{DLW} + \text{DSW}$$

Statistical analyses

Details on all statistical analyses used in this paper, including the statistical tests used, the number of replicates and precision measures, are indicated in the corresponding figure legends. Statistical analysis of replicate data was performed using Microsoft Excel 2017.

Reporting summary

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

Data availability

The sequencing raw data have been deposited in the NCBI database under accession code BioProject PRJNA1197171. The haploid reference genome is derived from a previously published study[5]. The assembled data have been deposited in the NCBI database under the BioProject accession code PRJNA1220045. Additionally, the data are available via Zenodo at https://doi.org/10.5281/zenodo.14118213 (ref. 85) and via Figshare at https://figshare.com/articles/dataset/Alfalfa/28426967 (ref. 86). Resequencing data used in this study were obtained from Zhang's research, and the relevant data have been provided in his published article[45]. The RNA sequence data from this study have been deposited in the NCBI database under accession code BioProject PRJNA1083622. The phenotypes used in GWAS and GS studies are available via Zenodo at https://doi.org/10.5281/zenodo.14869063 (ref. 87).

Code availability

All codes associated with this project are available via GitHub at https://github.com/hefei0609-afk/Alfalfa and via Zenodo at https://doi.org/10.5281/zenodo.14800545 (ref. 88).

References

45. Zhang, F. et al. Evolutionary genomics of climatic adaptation and resilience to climate change in alfalfa. *Mol. Plant* **17**, 867–883 (2024).
46. Li, H. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
47. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
48. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
49. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
50. Alexander, D. H. & Lange, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 246 (2011).
51. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
52. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).
53. Zhang, J. et al. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nat. Genet.* **50**, 1565–1573 (2018).
54. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859 (2005).
55. Tang, H. et al. An improved genome release (Version Mt4.0) for the model legume *Medicago truncatula*. *BMC Genomics* **15**, 312 (2014).
56. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
57. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
58. Guan, D. et al. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
59. Zhao, X. & Hao, W. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
60. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
61. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
62. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* **5**, 4.10.11–14.10.14 (2004).
63. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom. Bioinform.* **3**, lqaa108 (2021).
64. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
65. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
66. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
67. Su, W., Gu, X. & Peterson, T. TIR-Learner, a new ensemble method for TIR transposable element annotation, provides evidence for abundant new transposable elements in the maize genome. *Mol. Plant* **12**, 447–460 (2019).

68. Xiong, W. et al. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proc. Natl Acad. Sci. USA* **111**, 10263–10268 (2014).
69. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457 (2020).
70. Lavigne, R., Seto, D., Mahadevan, P., Ackermann, H.-W. & Kropinski, A. M. Unifying classical and molecular taxonomic classification: analysis of the Podoviridae using BLASTP-based tools. *Res. Microbiol.* **159**, 406–414 (2008).
71. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
72. Wang, D., Zhang, Y., Zhang, Z., Zhu, J. & Yu, J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics* **8**, 77–80 (2010).
73. Wang, D.-P., Wan, H.-L., Zhang, S. & Yu, J. γ-MYN: a new algorithm for estimating Ka and Ks with consideration of variable substitution rates. *Biol. Direct* **4**, 20 (2009).
74. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
75. Heller, D. & Vingron, M. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35**, 2907–2915 (2019).
76. Jiang, T. et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189 (2020).
77. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
78. Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
79. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
80. Zadeh, L. A. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets Syst.* **1**, 3–28 (1978).
81. VanRaden, P. M. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **91**, 4414–4423 (2008).
82. Cortes, C. & Vapnik, V. Support-vector networks. *Machine Leaning* **20**, 273–297 (1995).
83. Fu, C., Hernandez, T., Zhou, C. & Wang, Z.-Y. Alfalfa (*Medicago sativa* L.). *Methods Mol. Biol.* **1223**, 213–221 (2015).
84. Abràmoff, M. D., Magalhães, P. J. & Ram, S. J. Image processing with ImageJ. *Biophotonics Int.* **11**, 36–42 (2004).
85. He, F. Pan-genomic analysis highlights genes associated with agronomic traits and enhances genomics-assisted breeding in alfalfa. *Zenodo* https://doi.org/10.5281/zenodo.14118212 (2024).
86. He, F. Alfalfa. *Figshare* https://doi.org/10.6084/m9.figshare.28426967.v1 (2025).
87. Fei, H. Alfalfa. *Zenodo* https://doi.org/10.5281/zenodo.14869062 (2025).
88. Fei, H. Alfalfa pan-genome. *Zenodo* https://doi.org/10.5281/zenodo.14800544 (2025).

Acknowledgements

This work was supported by China Agriculture Research System of MOF and MARA (grant no. CARS-34 to Q.Y.), the Biological Breeding-National Science and Technology Major Project (grant no. 2022ZD04011 to R.L.), the Key Projects in Science and Technology of Inner Mongolia (grant no. 2021ZD0031 to R.L.) and Agricultural Science and Technology Innovation Program of CAAS (grant no. ASTIP-IAS14 to Q.Y.).

Author contributions

Q.Y., R.L. and X.Z. designed this project and coordinated the research activities. F.Z., J.K., H.L., L.C., Xianyang Li, M.L., X.W., X.J., B.S., M.X. and Y.L. collected and provided plant materials. F.Z., R.L. and X.Z. participated in the genome sequencing and resequencing. S.C., S.Q. and K.C. assembled the genomes. S.C., W.K., Q.Z., K.C. and S.Q. performed the gene annotation. S.C. and F.H. analyzed RNA-seq data. F.H. constructed the sequence and gene-based pan-genome. F.Z., S.C. and F.H. contributed to population GWAS analysis. Y.Z. performed functional verification. X.H., Xiao Li and T.Z. conducted a whole-genome selection analysis. F.H., S.C., X.Z., R.L. and Q.Y. interpreted the data and contributed to the manuscript writing.

Competing interests

The authors declare no competing interests.

Additional information

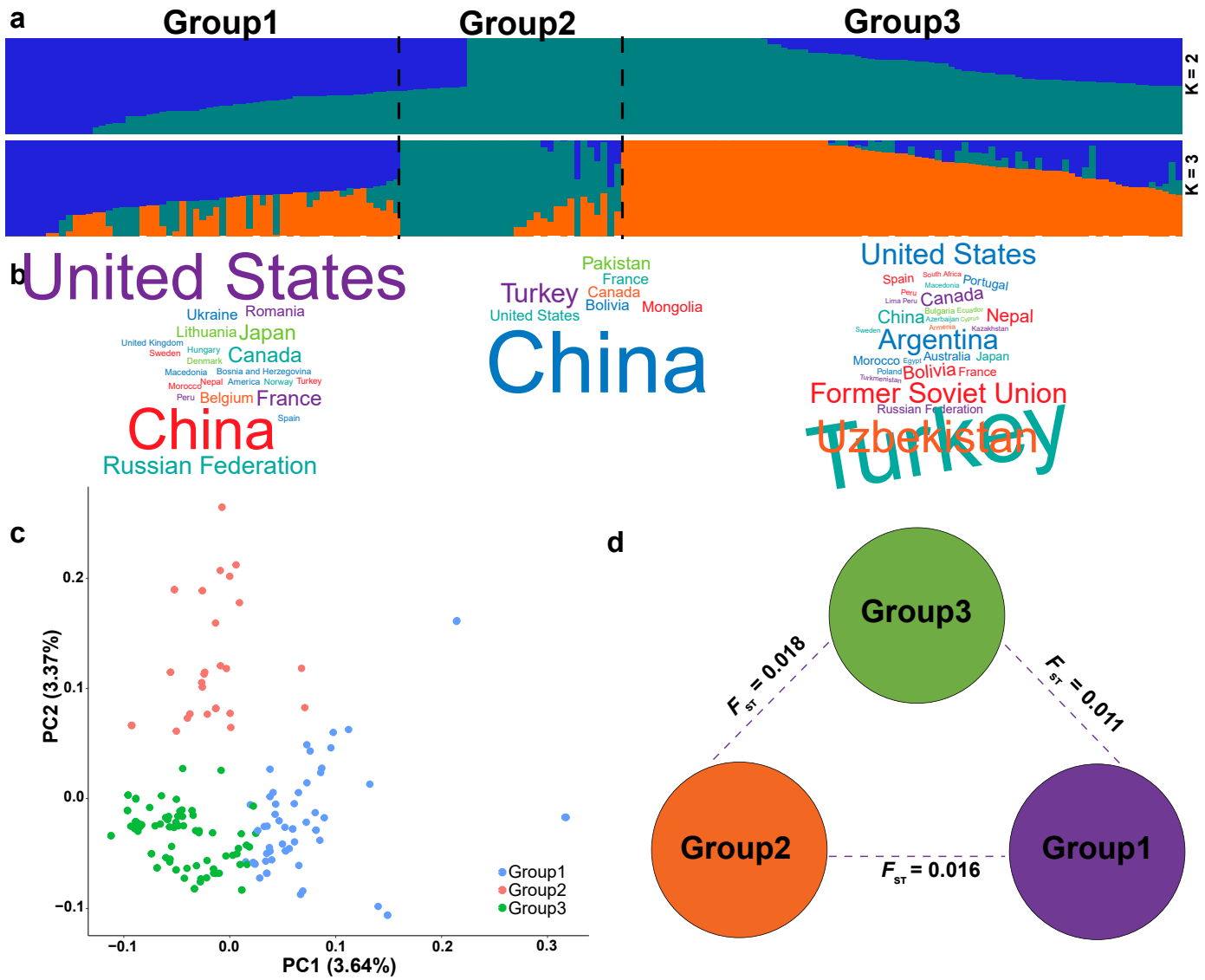
Extended data is available for this paper at https://doi.org/10.1038/s41588-025-02164-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-025-02164-8.

Correspondence and requests for materials should be addressed to Xingtan Zhang, Ruicai Long or Qingchuan Yang.

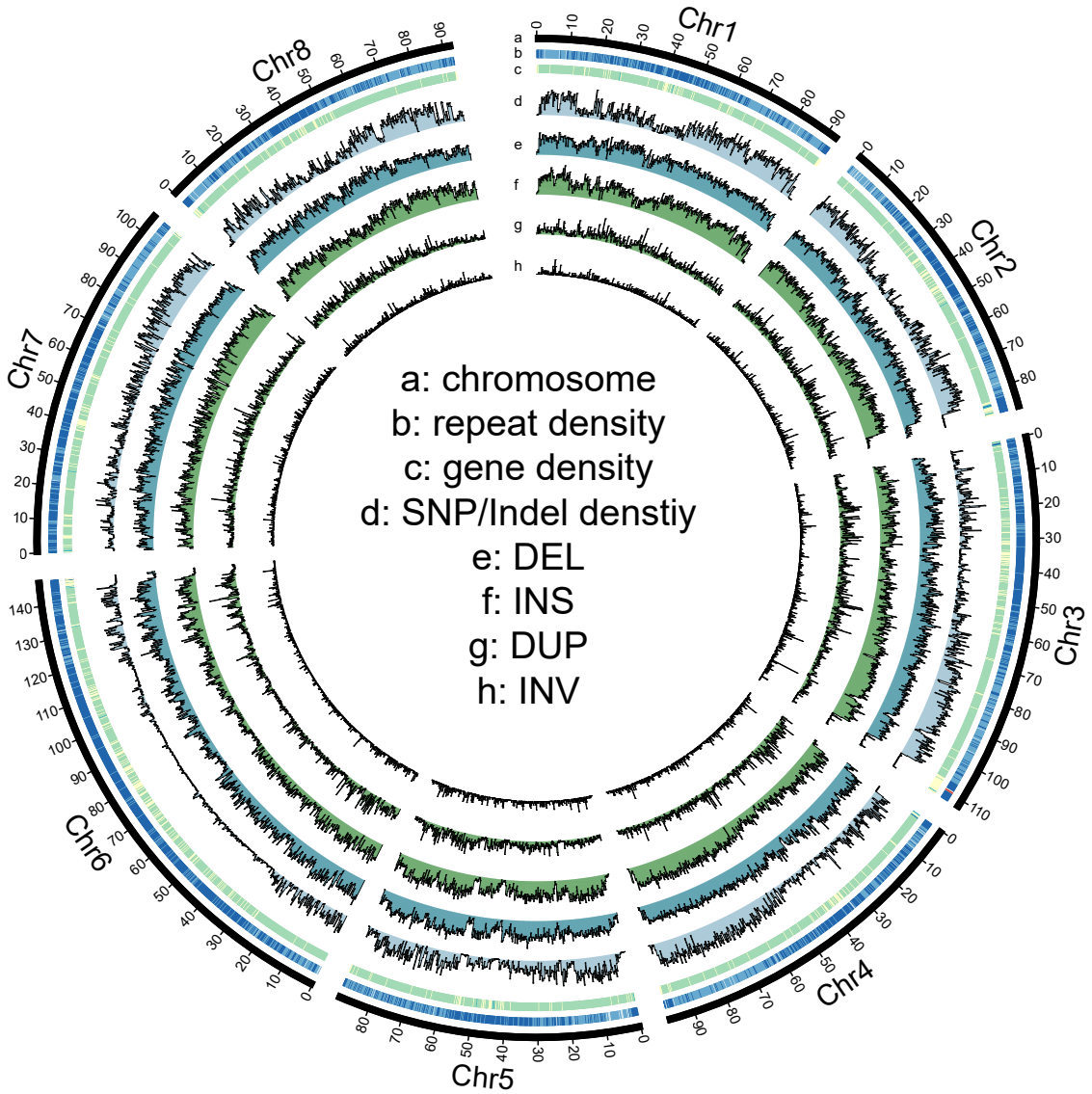
Peer review information *Nature Genetics* thanks Eric von Wettberg and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41588-025-02164-8



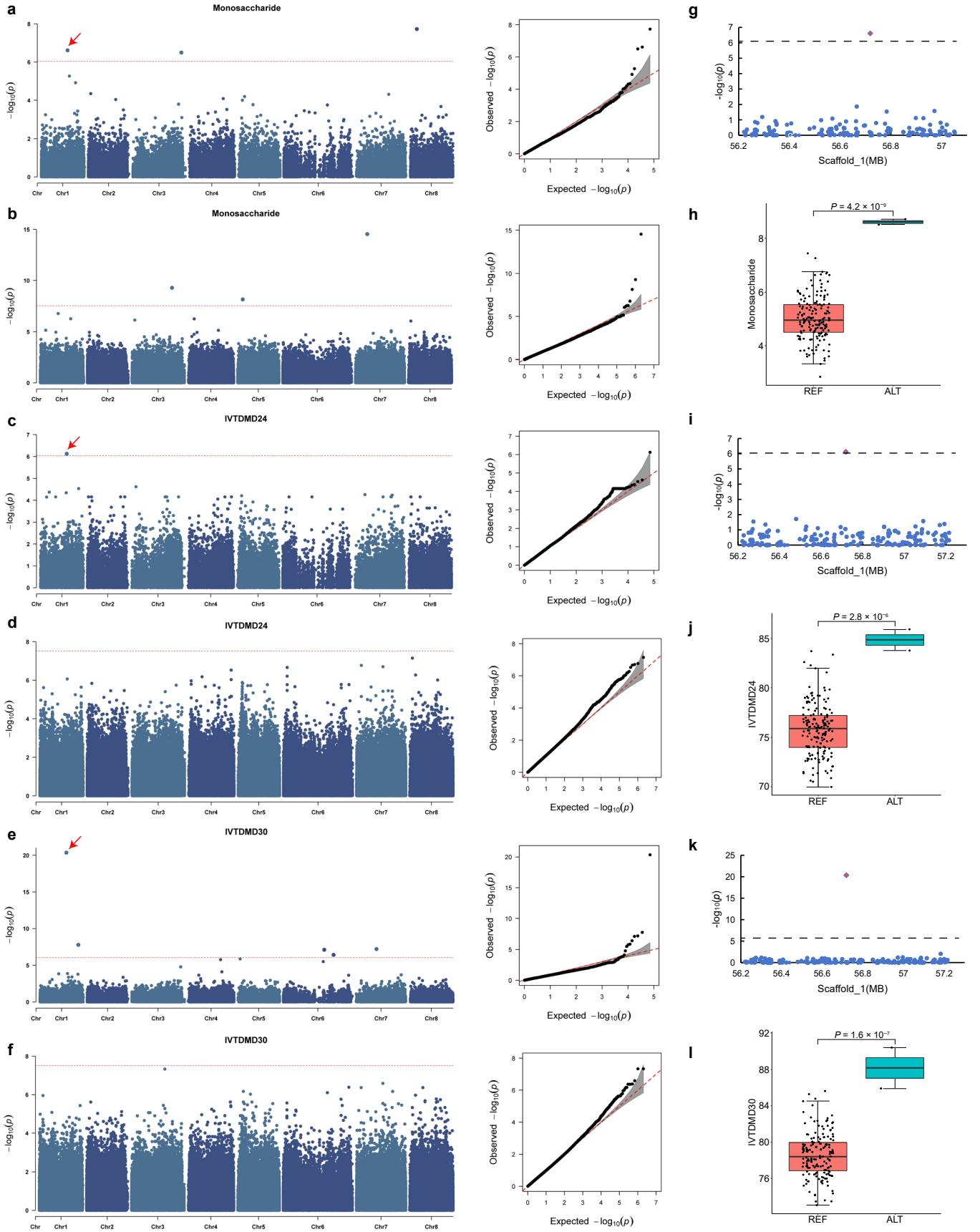
Extended Data Fig. 1 | Population structure and Fixation Index of the global alfalfa diversity panel. a. Population structure of the alfalfa panel was inferred by assuming three subpopulations (K). Each color represents a different subpopulation. **b**. Word cloud of the primary origin countries for alfalfa varieties in Group1, Group2, and Group3. Font size represents the relative proportion of varieties from each country. Group1 is predominantly from the United States, Group2 from China, and Group3 from Turkey, with contributions from other countries as well. **c**. The PCA scatter plot shows the distribution of PC1 and PC2, with different colors representing different groups (Group1, Group2, Group3). **d**. Fixation Index (F_{ST}) values among Group1, Group2, and Group3 alfalfa accessions.

https://doi.org/10.1038/s41588-025-02164-8



Extended Data Fig. 2 | The genome structure variations (SVs) between species of alfalfa. a, Chromosome. **b**–**h**, means the distribution of repeat density, gene density, SNP/Indel density, deletions, insertion, duplication and inversion.

https://doi.org/10.1038/s41588-025-02164-8

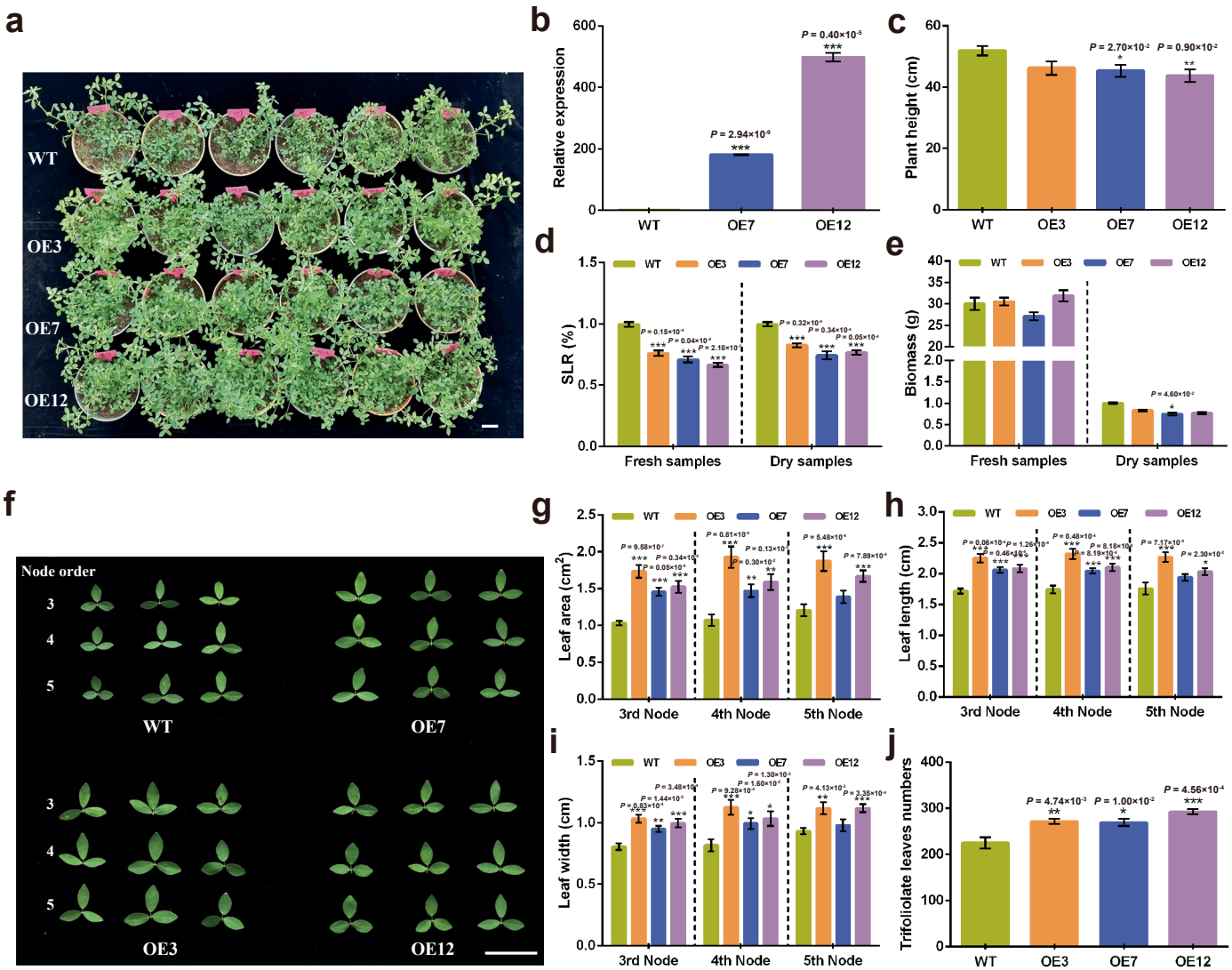


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41588-025-02164-8

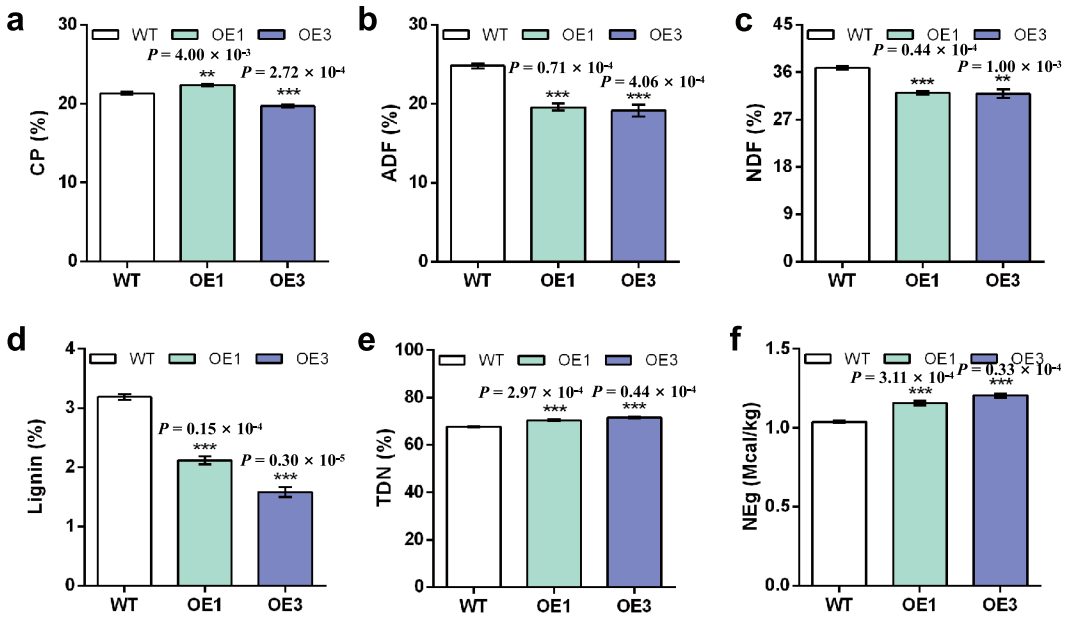
Extended Data Fig. 3 | Genome-wide association study (GWAS) for monosaccharide content, In Vitro True Dry Matter Degradability at 24 h (IVTDMD24), and In Vitro True Dry Matter Degradability at 30 h (IVTDMD30). a, c, e, present the Manhattan and QQ plots of the GWAS results for monosaccharide, IVTDMD24, and IVTDMD30, respectively, using structural variation (SV) markers. **b, d, f**, show the Manhattan and QQ plots for the same traits using single nucleotide polymorphism (SNP) markers. The red dashed line indicates the Bonferroni-corrected genome-wide significance threshold ($\alpha = 0.05/n$, where 'n' is the total number of independent SNPs and effective SVs). **g, i, k**, depict scatter plots of the peak structural variations in chromosome 1 for the three traits, with the horizontal line marking the Bonferroni-corrected genome-wide significance threshold. **h, j, l**, display boxplots of the three traits across different accessions, categorized by the alleles they carry. The sample sizes for the REF and ALT groups are 171 and 5, respectively. In boxplots, the 25% and 75% quartiles are shown as lower and upper edges of boxes, respectively, and central lines denote the median. The whiskers extend to 1.5 times the inter-quartile range. *P*-values were computed from two-tailed Student's *t*-test.

https://doi.org/10.1038/s41588-025-02164-8



Extended Data Fig. 4 | Impact of *MsGA3ox1* overexpression on alfalfa morphology traits. a, Comparison between WT alfalfa plants and overexpression lines (OE3, OE7, and OE12). **b-e**, Quantitative measurements of *MsGA3ox1* expression levels, plant height, SLR, and biomass. **f**, Photographs of leaves from WT, OE3, OE7, and OE12 lines at the 3rd, 4th, and 5th stem nodes. **g-i**, Comparative assessments of leaf area, leaf length, and leaf width between WT and *MsGA3ox1* overexpression lines as shown in f. **j**, Comparison of WT and *MsGA3ox1* overexpression lines in the number of trifoliolate leaves. The scale bar represents 5 cm. Asterisks denote statistical significance with '*' '**' and '***' indicating $P < 0.05$, $P < 0.01$ and $P < 0.001$, respectively. Data are presented as means ± SEM, with three independent experimental replicates for panel b, six independent experimental replicates for panels c, d, e, and j, and nine independent experimental replicates for panels g, h, and i. The control group (WT) is the Zhongmu No.1 variety of *Medicago sativa* L.

https://doi.org/10.1038/s41588-025-02164-8



Extended Data Fig. 5 | Phenotypic characterization of alfalfa quality traits in *MsGA3ox* overexpression lines. The bar graphs depict a comparative analysis of crude protein (CP) (**a**), acid detergent fiber (ADF) (**b**), neutral detergent fiber (NDF) (**c**), lignin content (**d**), total digestible nutrients (TDN) (**e**), and net energy for gain (NEg) (**f**) between WT and overexpressed lines OE1 and OE3. Asterisks denote levels of statistical significance compared to WT (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$). Data are presented as means ± SEM, with four biological replicates per group. The control group (WT) is the Zhongmu No.1 variety of *Medicago sativa* L.

nature portfolio

Corresponding author(s): Qingchuan Yang

Last updated by author(s): Nov 13, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The sequencing reads were undertaken with PacBio Sequel II and BGISeq T7 platforms.
Data analysis	We used publicly available and appropriately cited software as described. No commercial software or code was used in this study. The list of software is as follows: BWA (v0.7.17- ｢1188),samtools (vl.13), vcftools (v0.1.16), fasttree (v2.l.11), ADMIXTURE (vl.3.0),, Juicer (vl.5), ALLHiC (https://github.com/tangerzhang/ALLHiC/wiki/ALLHiC), GMAP (v2013-10-28), MCScanX (https://github.com/tanghaibao/mcscan), hifiasm (https://github.com/chhylp123/hifiasm; v0.19.8-r603), algorithm (vl.0), purge_dups (vl.01), BUSCO (v5.0), RagTag (v1.1.1), RepeatModeler (v1.0.11 and v2.0), LTR_FINDER (vl.0.7), LTR_harvest (vl.5.10), LTR_retriever (vl.6), RepeatMasker (vl.332), BRAKER (v2.l.6), AUGUSTUS (v3.3.0), GeneMark-ES/ET (v3.67), PFAM (https://pfam-legacy.xfam.org/), InterProScan5 (http://lilab2.sysu.edu.cn/Tools/pfa/iprscan5/help/), EggNOG-mapper (http://eggnog-mapper.embl.de/), EDTA (v2.0.0), TIR-Learner (v2.5), HelitronScammer (vl.1), OrthoFinder (v2.5.2), KaKs (v19), NGLMR (v0.2.7), SVIM (v2.0.0), cuteSV (v2.0.3), Sniffles (vl.0.12), SURVIVOR (vl.0.7), rMVP (vl.0.6), BCFTools (vl.13), Variant Graph (vg) (vl.34.0), Image J (1.53k) , CMplot package (v4.5.1),Khaper (v1.0), R (v4.3.1), Pfam(v35).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The sequencing raw data have been deposited in the NCBI database under accession code BioProject PRJNA1197171. The haploid reference genome is derived from a previously published study5. The assembled data have been deposited in the NCBI database under the BioProject accession code PRJNA1220045. Additionally, the data are available on Zenodo at the following DOI: https://doi.org/10.5281/zenodo.14118213. Resequencing data used in this study was obtained from Zhang's research, and the relevant data has been provided in his published article45. The RNA sequence data from this study have been deposited in the NCBI database under accession code BioProject PRJNA1083622. The phenotypes used in GWAS and GS studies have been deposited in https://doi.org/10.5281/zenodo.14869063.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	not applicable
Reporting on race, ethnicity, or other socially relevant groupings	not applicable
Population characteristics	not applicable
Recruitment	not applicable
Ethics oversight	not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We selected 24 representative alfalfa accessions, including 2 wild materials, 2 accessions with unclear improvement status, 11 landrace accessions, and 9 cultivated accessions. The selection logic was based on phylogenetic relationships, geographic distribution, breeding and/or research contributions, and subgroup distribution to ensure that they represent the genetic diversity of alfalfa.
Data exclusions	No data were excluded.
Replication	In the qRT-PCR, plant height, branch number, stem-leaf ratio, and biomass experiments for WT, OE1, and OE3, four biological replicates were used per group. For leaf length, leaf width, and leaf area experiments for WT, OE1, and OE3, nine biological replicates were used per group. In the second batch of experiments, OE3, OE7, and OE12 were tested with six biological replicates. All replications were successful and were used.
Randomization	The samples in this study were randomly sampled.
Blinding	Blinding is not necessary for genome sequencing and assembly, since the investigators know which alfalfa accessionsthey were handling.The investigators were blinded to group allocation during collecting data from WT and transgenic lines

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☐	☒ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes
☒	☐ Public health
☒	☐ National security
☒	☐ Crops and/or livestock
☒	☐ Ecosystems
☒	☐ Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes
☒	☐ Demonstrate how to render a vaccine ineffective
☒	☐ Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	☐ Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	☐ Increase transmissibility of a pathogen
☒	☐ Alter the host range of a pathogen
☒	☐ Enable evasion of diagnostic/detection modalities
☒	☐ Enable the weaponization of a biological agent or toxin
☒	☐ Any other potentially harmful combination of experiments and agents

Plants

Seed stocks	The germplasm resources used in this study were obtained from the Medium Term Library of the National Grass Seed Resources of China and the U.S. National Plant Germplasm System (https://npgsweb.ars-grin.gov/gringlobal/search). Catalogue numbers and other detailed information for each germplasm resource can be found in Supplementary Table 1. No plant specimens were collected from the field in this study.
Novel plant genotypes	In this study, transgenic alfalfa was generated using an Agrobacterium-mediated transformation method. The full-length coding sequence of MsGA3ox was obtained by PCR amplification from Medicago sativa L. cv. Zhongmu-1. To construct the overexpression vector, the MsGA3ox plasmid was amplified using primers MsGA3ox-1300-F and MsGA3ox-1300-R, and then recombined with the Super1300 expression vector containing the SmaI restriction site via homologous recombination. The plasmid with the target gene for overexpression was subsequently transformed into Agrobacterium EHA105 competent cells, and the desired gene was introduced into alfalfa following Fu et al.'s protocol, completing the genetic transformation process. After approximately seven months of callus culture, bud culture, and root culture, regenerants were obtained.
Authentication	Following hygromycin (5mg/L) resistance screening and DNA and RNA level identification, a total of four independent MsGA3ox overexpressing (OE) alfalfa lines were selected for subsequent experiments.