ORIGINAL RESEARCH

Genome Assembly of Alfalfa Cultivar Zhongmu-4 and Identification of SNPs Associated with Agronomic Traits

Ruicai Long [1,#], Fan Zhang [1,2,#], Zhiwu Zhang [2], Mingna Li [1], Lin Chen [1], Xue Wang [1], Wenwen Liu [1], Tiejun Zhang [3], Long-Xi Yu [4], Fei He [1], Xueqian Jiang [1], Xijiang Yang [1], Changfu Yang [1], Zhen Wang [1,*], Junmei Kang [1,*], Qingchuan Yang [1,*]

[1] Institute of Animal Sciences, Chinese Academy of Agricultural Sciences, Beijing 100193, China
[2] Department of Crop and Soil Sciences, Washington State University, Pullman, WA 99163, USA
[3] School of Grassland Science, Beijing Forestry University, Beijing 100083, China
[4] United States Department of Agriculture-Agricultural Research Service, Plant and Germplasm Introduction and Testing Research, Prosser, WA 99350, USA

Abstract Alfalfa (*Medicago sativa* L.) is the most important legume forage crop worldwide with high nutritional value and yield. For a long time, the breeding of alfalfa was hampered by lacking reliable information on the **autotetraploid** genome and molecular markers linked to important agronomic traits. We herein reported the *de novo* assembly of the allele-aware chromosome-level genome of Zhongmu-4, a cultivar widely cultivated in China, and a comprehensive database of genomic variations based on **resequencing** of 220 germplasms. Approximate 2.74 Gb contigs (N50 of 2.06 Mb), accounting for 88.39% of the estimated genome, were assembled, and 2.56 Gb contigs were anchored to 32 pseudo-chromosomes. A total of 34,922 allelic genes were identified from the allele-aware genome. We observed the expansion of gene families, especially those related to the nitrogen metabolism, and the increase of repetitive elements including transposable elements, which probably resulted in the increase of Zhongmu-4 genome compared with *Medicago truncatula*. Population structure analysis revealed that the accessions from Asia and South America had relatively lower genetic diversity than those from Europe, suggesting that geography may influence

* Corresponding authors.
  E-mail: wangzhen@caas.cn (Wang Z), kangjunmei@caas.cn (Kang J), yangqingchuan@caas.cn (Yang Q).
\# Equal contribution.

alfalfa genetic divergence during local adaption. **Genome-wide association studies** identified 101 single nucleotide polymorphisms (SNPs) associated with 27 agronomic traits. Two candidate genes were predicted to be correlated with fall dormancy and salt response. We believe that the allele-aware chromosome-level genome sequence of Zhongmu-4 combined with the resequencing data of the diverse alfalfa germplasms will facilitate genetic research and genomics-assisted breeding in variety improvement of alfalfa.

Introduction

Alfalfa (*Medicago sativa* L.), one of the most widely cultivated perennial forage crops worldwide, has been considered as the most valuable forage species because of its high yield and nutritional value [1]. Alfalfa is also relatively tolerant to stress and able to fix nitrogen [2,3]. The average annual alfalfa hay production in the United States has exceeded 53 million tons in the last three years (Crop Production Summary 2020, United States Department of Agriculture, https://www.ams.usda.gov/market-news/hay-reports). In China, alfalfa hay production in 2019 reached 16.94 million tons from a planting area of approximate 2.32 million ha [4]. With a rapidly increasing demand for milk and ruminant animal production in the developing countries including China, the shortage of high-quality forage crops, particularly alfalfa, has become a limiting factor for the sustainable production of high-quality animals.

Cultivated alfalfa is a cross-pollinated perennial autotetraploid (2n = 4× = 32) species. The complex genetic feature of alfalfa makes it difficult to develop new cultivars using an efficient breeding strategy. Consequently, recurrent phenotypic selection, a time-consuming and labor-intensive method, has been used predominantly for alfalfa breeding. Marker-assisted breeding, an efficient alternative to the traditional phenotypic selection, has been applied and greatly helped in revealing the genetic architecture of the agronomic traits [2,5]. Several candidate single nucleotide polymorphism (SNP) markers have been identified for quantitative traits, including biomass [6–8], forage quality [9–12], and drought/salt tolerance [13–15]. Quantitative trait loci (QTLs) associated with yield [16–18] and fall dormancy [19,20] have also been identified. Additionally, a genomic selection-related study based on the alfalfa biomass yield has been conducted [21]. However, the reference genome used in most of these studies is from *Medicago truncatula*, which is a close relative of alfalfa [22]. Although *M. sativa* and *M. truncatula* are in the same genus, there are some distinct features (*e.g.*, ploidy, overwintering rate, and stress tolerance). Furthermore, some loci of alfalfa are unable to be mapped to the *M. truncatula* genome, reflecting the divergence of alfalfa from its relatives [23,24]. A highly resolved genome is a valuable resource for improving the important agronomic traits of cultivated alfalfa via marker-based breeding.

Assembling chromosome-level genome sequences for autopolyploid or highly heterozygous genomes has historically been a monumental challenge. In 2018, the first allele-aware chromosome-level genome of the autopolyploid *Saccharum spontaneum* was assembled using the ALLHiC algorithm developed by Zhang and his colleagues [25,26]. Two recently published draft genomes of the autopolyploid alfalfa, Xinjiangdaye and Zhongmu-1, provide valuable information for related studies of alfalfa [27,28]. The Xinjiangdaye genome was assembled at the allele-aware chromosome level using the ALLHiC algorithm, whereas the Zhongmu-1 genome was merely assembled at the haploid chromosome level. Compared with several other plants such as maize and rice, the genomic information of cultivated alfalfa is still limited. In this study, we assembled the allele-aware chromosome-level genome sequence of the Chinese alfalfa cultivar Zhongmu-4, which is one of the most common alfalfa cultivars grown in northern China because of its high yield and salt tolerance. Zhongmu-4 alfalfa was bred by us and released in 2011. Advantages of this cultivar include its greater root system, large leaf area, high nutritional value (about 20% crude protein content), fast regeneration, high yield, and salt tolerance. Zhongmu-4 alfalfa is widely cultivated in the Huang-Huai-Hai region of China. The mean annual hay yield can reach 15.0–17.4 t/ha, which is about 10% greater than that of Zhongmu-1. By the end of 2020, the total planting area of Zhongmu-4 has exceeded 200,000 ha. Moreover, in the present study, we evaluated 93 agronomic traits of 220 alfalfa accessions from a global germplasm resource. On the basis of our assembled genome sequence and genotyping by whole-genome sequencing, 111,075 SNP markers were screened. Several SNPs associated with agronomic traits (*e.g.*, salt tolerance and fall dormancy) were identified by genome-wide association studies (GWAS).

Results

Allele-aware genome assembly and annotation

Flow cytometry analysis revealed that the DNA peak ratio of Zhongmu-4 to *M. truncatula* ecotype Jemalong A17, a close relative of alfalfa with an estimated haploid genome size of 425 Mb (1C) [22], was about 3.61 (Figure S1), indicating that the tetraploid genome of Zhongmu-4 was approximately 3068 Mb (2C) (Table S1). The estimation was close to the genome size (3371 Mb) obtained based on the 2C value (3.44 pg) in the Plant DNA C-values database. Using the 168 Gb Illumina 150 bp paired-end genome sequencing reads, the Zhongmu-4 genome was surveyed and the tetraploid genome size was estimated as 2962 Mb by KmerGenie according to the K-mer frequency ($K = 19$) (Figure S2). Based on the aforementioned estimates, the estimated Zhongmu-4 genome size was set at 3.1 Gb for further analyses.

The 262 Gb continuous long reads (CLRs) generated by the PacBio Sequel sequencing system (Table S2) were used for the self-correction and assembly by Canu. Approximately 2.74 Gb contigs, accounting for 88.39% of the estimated genome size, were assembled with contig N50 of 2.06 Mb and GC content of 34.2% (Table S3). The assembled contig sequences were corrected and polished using 168 Gb Illumina paired-end sequences, and then preliminarily assembled using 3D-DNA, yielding 49,967 corrected contigs (Table S3). The corrected and polished contigs were imported to the polyploid genome

assembly program ALLHiC to build an allele-aware chromosome-scale genome using high-through chromosome conformation capture (Hi-C) paired-end reads. The final assembled genome included 2.56 Gb contigs anchored to 32 pseudo-chromosomes (super-scaffolds), which represented eight groups of homologous chromosomes with four sets of monoploid chromosomes, and 0.18 Gb unanchored contigs. The assembly quality was assessed according to the synteny analysis, Hi-C contact matrix, benchmarking universal single-copy orthologs (BUSCO), and the long terminal repeat (LTR) assembly index (LAI). The Hi-C linkage plot indicated that the chromosome groups were distinct (Figure 1A; Table S4). The sequence alignment and synteny analysis revealed high collinearity between the individual Zhongmu-4 subgenome and the Jemalong A17 genome (Figure 1B and C, Figure S3). Approximately 98.4% of BUSCOs (1588 out of 1614) were detected in the assembled genome (Table S5). The LAI of the assembled genome was about 13.85, indicating that Zhongmu-4 genome assembly reached the Reference level of Ou's classification system [29]. Thus, these results indicate the high quality of the assembled Zhongmu-4 genome.

A total of 146,704 protein-coding genes were identified in the assembled allele-aware genome, and 136,467 genes (93.02%) were functionally annotated according to the Inter-Pro, eggNOG, and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases. The gene density in the 32 monoploid chromosomes is presented in Figure 2 and Table S6. Using the first set of subgenomes as the reference, 34,922 allelic genes were identified in the four sets of allele-aware chromosomes (10,704 four-allele genes, 9526 three-allele genes, 7129 two-allele genes, and 7563 one-allele genes).

Repetitive elements drive genome expansion of alfalfa

A total of 1558 Mb repetitive sequences were identified, accounting for 56.80% of the assembled Zhongmu-4 genome (Figure 2), of which 49.28% were LTR transposable elements (TEs), including Ty1-*Copia* superfamily retrotransposons (10.93%), Ty3-*Gypsy* superfamily retrotransposons (18.33%), and unclassified LTR elements (20.02%). The remaining repetitive sequences were mainly long or short interspersed nuclear element segments (LINEs or SINEs), DNA elements, and
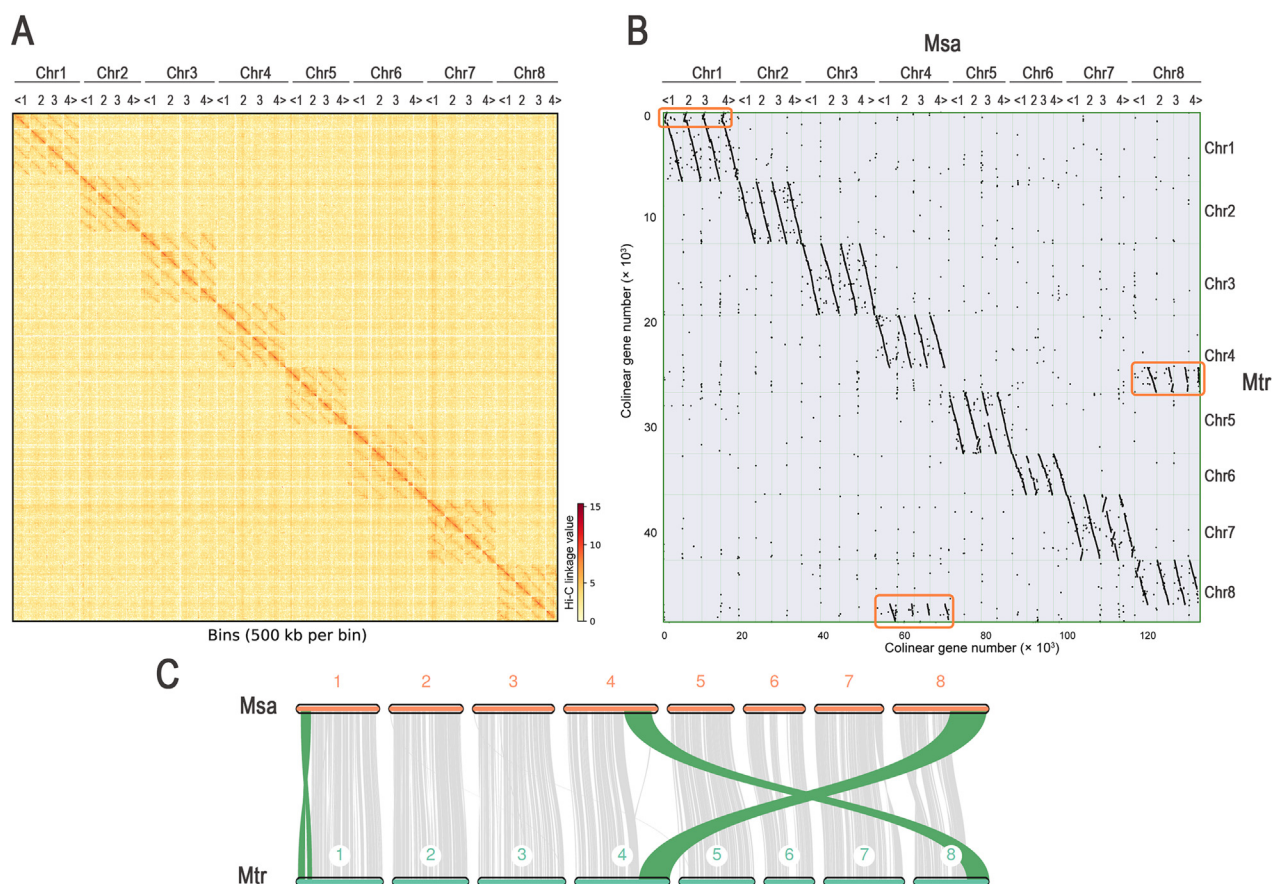

Figure 1 Hi-C linkage and synteny analyses

A. Hi-C linkage plot of 32 chromosomes in the Zhongmu-4 genome. **B.** Alignment and synteny analysis of four subgenomes of *Medicago sativa* cultivar Zhongmu-4 and the genome of the *M. truncatula* ecotype Jemalong A17. The translocation between Chr4 and Chr8 and the inversion in Chr1 were labeled by orange boxes. **C.** Synteny linkage plot of the first set of monoploid chromosomes of Zhongmu-4 and Jemalong A17. Hi-C, high-through chromosome conformation capture; Chr, chromosome; Msa, *M. sativa* cultivar Zhongmu-4; Mtr, *M. truncatula* ecotype Jemalong A17.
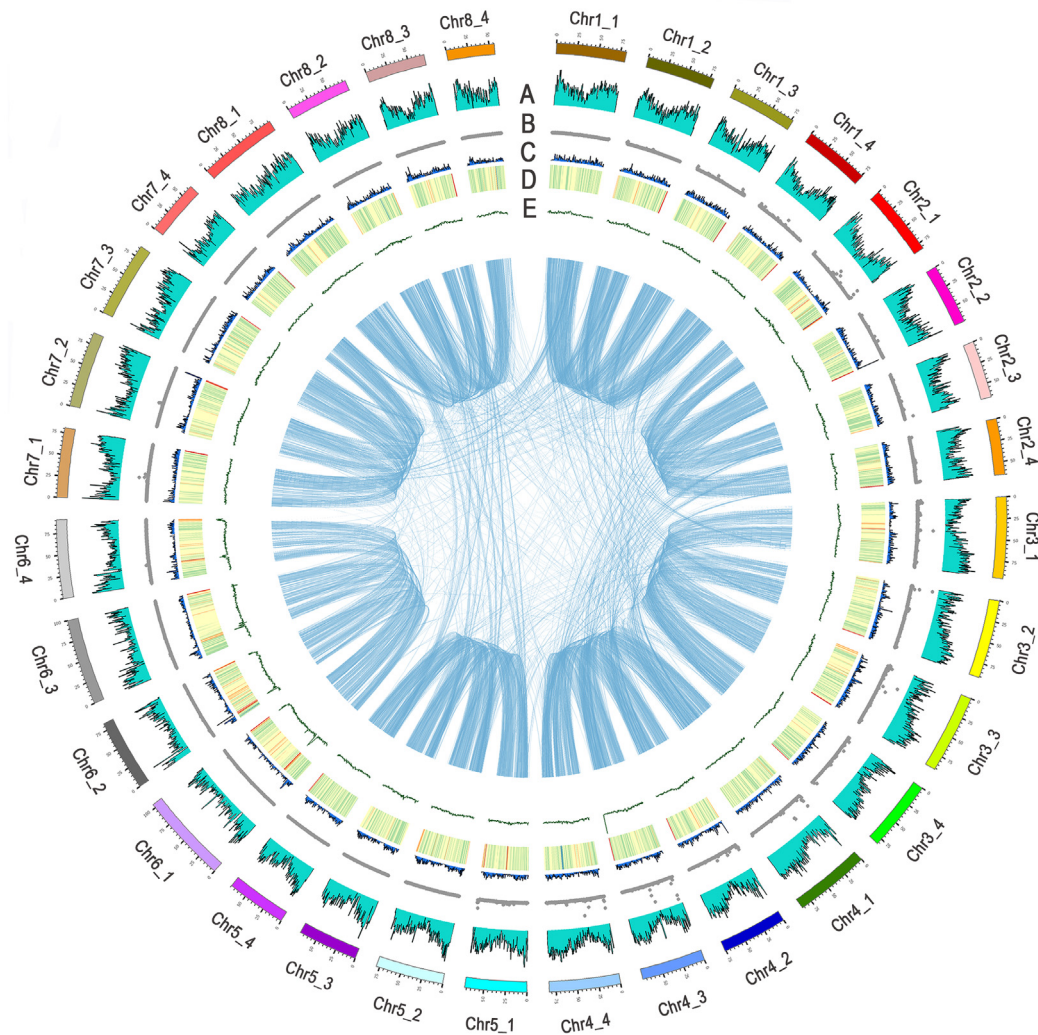
Figure 2 Distribution of Zhongmu-4 genomic features
A. Density of annotated genes. **B.** Density of filtered SNPs in the GWAS population. **C.** Density of syntenic paralogous gene pairs with a Ka/Ks ratio > 1. **D.** Density of repetitive sequences. **E.** GC content density. Synteny blocks are linked in the center. The outermost bars represent the length (Mb) of chromosomes. SNP, single nucleotide polymorphism; GWAS, genome-wide association study; Ka, the number of nonsynonymous substitutions per nonsynonymous site; Ks, the number of synonymous substitutions per synonymous site.

simple repeats. In contrast, *M. truncatula* contained 181 Mb repetitive sequences (43.94% of the genome), including 76 Mb LTR TEs, accounting for 18.36% of the genome. Thus, the proportion of repetitive sequences, especially LTR elements, was larger in Zhongmu-4 than in *M. truncatula*. These results suggest that the accumulation of LTR is probably one of the main reasons for the expansion of alfalfa genome. Regarding individual chromosomes, the group 6 chromosomes (Chr6_1 to Chr6_4) had a larger proportion of repetitive sequences than the other chromosomes (Figures 2, Figure S4A). These observations may explain, at least to some extent, why the assembly quality of chromosome 6 was relatively low with fewer protein-coding genes detected compared with the other chromosomes (Table S6). The distribution of LTR TE insertion time indicated that an LTR TE burst occurred in the Zhongmu-4 genome within the last one million years (Figure S4B). The LTRs in the group 6 chromosomes

were more active than those in the whole genome (Figure S4C and D).

Gene family evolution and whole-genome duplication

A phylogenetic tree was constructed using the 28 single-copy orthologous genes identified in *M. sativa* (the first subgenome of Zhongmu-4 was used), *Arabidopsis thaliana*, *Populus trichocarpa,* and 12 legume species (**Figure 3**A). The time of the divergence between *A. thaliana* and *M. sativa*, 106 million years ago (MYA), was used as a reference. *M. sativa* was more closely related to the legume species than to *A. thaliana* or *P. trichocarpa* (Figure 3A). Moreover, it diverged from *M. truncatula* approximately 5.3 MYA. The gene family size changed in the 15 selected species during evolution, with 3377 expanded gene families and 5626 contracted gene families in the first subgenome of Zhongmu-4 (Figure 3A). Gene
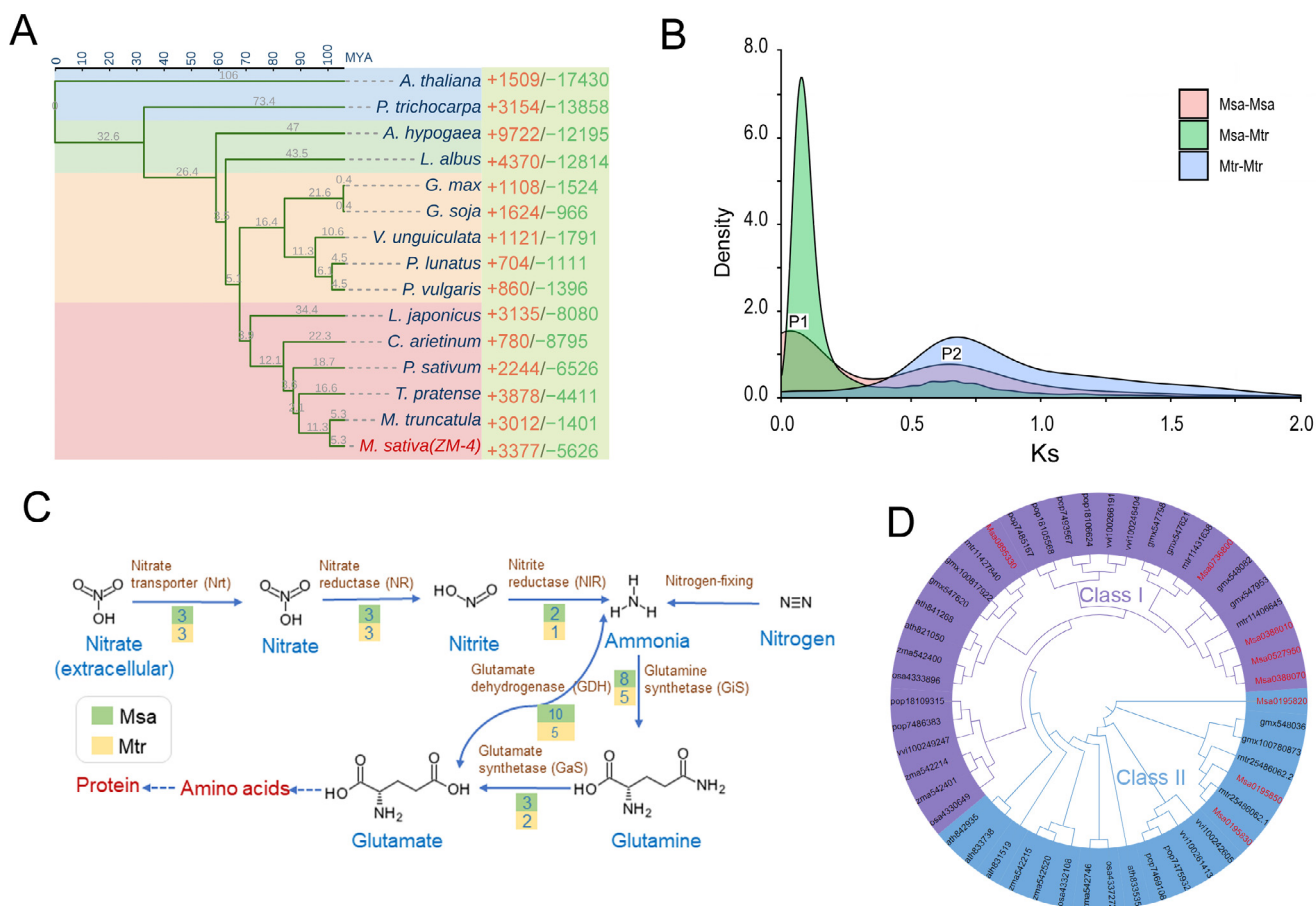
Figure 3　Comparative genome analysis and identification of nitrogen metabolism-related genes
A. Phylogenetic tree of Zhongmu-4 and 14 dicotyledonous species. " + " and "−" indicate expanded (red) and contracted (green) gene families, respectively. **B.** Density distribution of Ks values of the syntenic paralogous genes in/between the first subgenome of Zhongmu-4 and the genome of Jemalong A17. **C.** Key enzymes in the nitrogen metabolism pathway. The number in each colored block indicated the number of enzyme-coding genes. **D.** Phylogenetic tree of glutamine synthetases from Zhongmu-4 (Msa), Jemalong A17 (mtr), *G. max* (gmx), *A. thaliana* (ath), *Z. mays* (zma), *O. sativa* (osa), *V. vinifera* (vvi), and *P. trichocarpa* (pop). The eight glutamine synthetases from Zhongmu-4 were marked in red. The NCBI-Gene IDs were used to represent the glutamine synthetases from other seven species. MYA, million years ago.

Ontology (GO) annotation results showed that both the expanded and contracted genes were involved in 25 common biological process terms, predominantly in metabolic process and cellular process. Interestingly, some expanded genes were annotated in two biological process terms, nitrogen utilization and carbohydrate utilization (Figure S5). According to the gene numbers of the expanded and contracted families, PIF1-like helicase, TIR-NBS-LRR disease resistance protein, and heat shock 70 protein are the most notable expanded gene families, while MADS-box transcription factor, transcription termination factor, and F-box protein interaction domain protein are the most notable contracted gene families.

Synteny block analysis is often conducted to assess the assembly quality and investigate the evolution of related species. Following the protein sequence alignment, the syntenic blocks revealed that the four Zhongmu-4 subgenomes had strong collinear relationships with the Jemalong A17 genome. A translocation between Chr4 and Chr8 and a consistent inversion in Chr1 were detected (Figure 1B and C, Figure S3). The detection of the translocation and inversion was consistent with the previous study [27].

We analyzed the number of synonymous substitutions per synonymous site (Ks), the number of nonsynonymous substitutions per nonsynonymous site (Ka), and the Ka/Ks ratio in Zhongmu-4. The Ks value densities of Zhongmu-4 and Jemalong A17 revealed one overlapping peak P2 (Msa-Msa: 0.65), implying that they experienced a common ancient whole-genome duplication event before their divergence (Figure 3B, Figure S6). The Ks peak P1 (Msa-Msa: 0.04) of Zhongmu-4 was mainly caused by tandem repeat, segmental duplication, and high-frequency recombination of some allelic chromosome regions. The distribution of genes with Ka/Ks > 1 (between syntenic paralogous gene pairs) is presented in Figure 2. There was no significant difference among the allelic chromosomes.

Nitrogen metabolism-related gene families expand in Zhongmu-4

A BLASTP search identified genes encoding six key enzymes [*e.g.*, glutamine synthetase (GiS)] in the nitrogen metabolism pathway in the first subgenome of Zhongmu-4. The genes

encoding these enzymes identified in Zhongmu-4 and seven other species are listed in Table S7. The number of genes encoding nitrite reductase (NIR), GiS, glutamate dehydrogenase (GDH), or glutamate synthetase (GaS) was significantly larger in Zhongmu-4 than in Jemalong A17 (Figure 3C), implying that these gene families expanded after the divergence from Jemalong A17. The genes encoding GDH or GiS were more abundant in Zhongmu-4 than in six other analyzed species (Table S7). A phylogenetic tree constructed based on the protein sequence of GiS from Zhongmu-4 and seven other
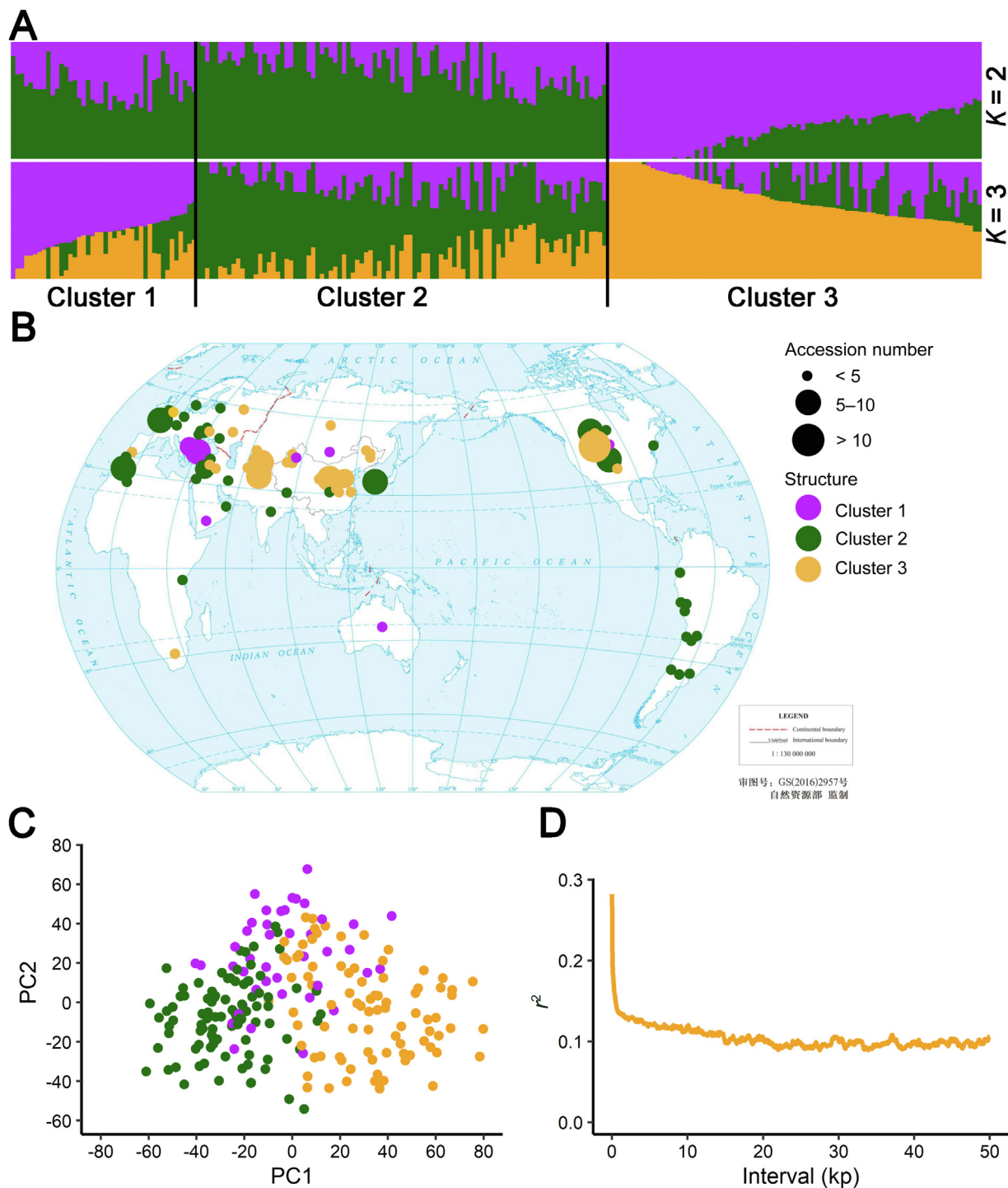
Figure 4　Analyses of the population structure and LD

A. Population structure cluster plot. **B.** Global distribution of the 220 accessions included in the population cluster plot. The base world map is from China Standard Map Service (http://bzdt.ch.mnr.gov.cn/). **C.** PCA involving PC1 and PC2 pairwise scatter plots. **D.** LD decay within 50 kb. PCA, principal component analysis; PC, principal component; LD, linkage disequilibrium.

species revealed that the GiS proteins were obviously divided into two clusters ([Figure 3](#)D).

SNP identification and population structure differentiation

To explore SNPs related to alfalfa agronomic traits, 220 accessions collected from six continents were genotyped by whole-genome sequencing. About 29.6 million SNPs detected using BWA-SAMtools-VarScan pipeline were filtered according to the following criteria: missing value > 10%, minimum mean read depth > 5, and minor allele frequency (MAF) > 0.05.

On the basis of 111,075 remaining SNP markers, the subgroup population structure was estimated using Admixture, with a K value of 2–10. The 220 accessions were divided into three clusters according to the cross-validation (CV) error ([Figure 4](#)A, Figure S7). Cluster 1 primarily comprised accessions from Europe, Cluster 2 included accessions with diverse backgrounds, and Cluster 3 mainly consisted of accessions from Asia ([Figure 4](#)B). A principal component analysis (PCA) showed similar results ([Figure 4](#)C), and the variations explained by the top three principal components (PCs) were 3.5%, 1.8%, and 1.5%, respectively, reflecting the relatively weak population structure of the alfalfa accessions. Linkage disequilibrium
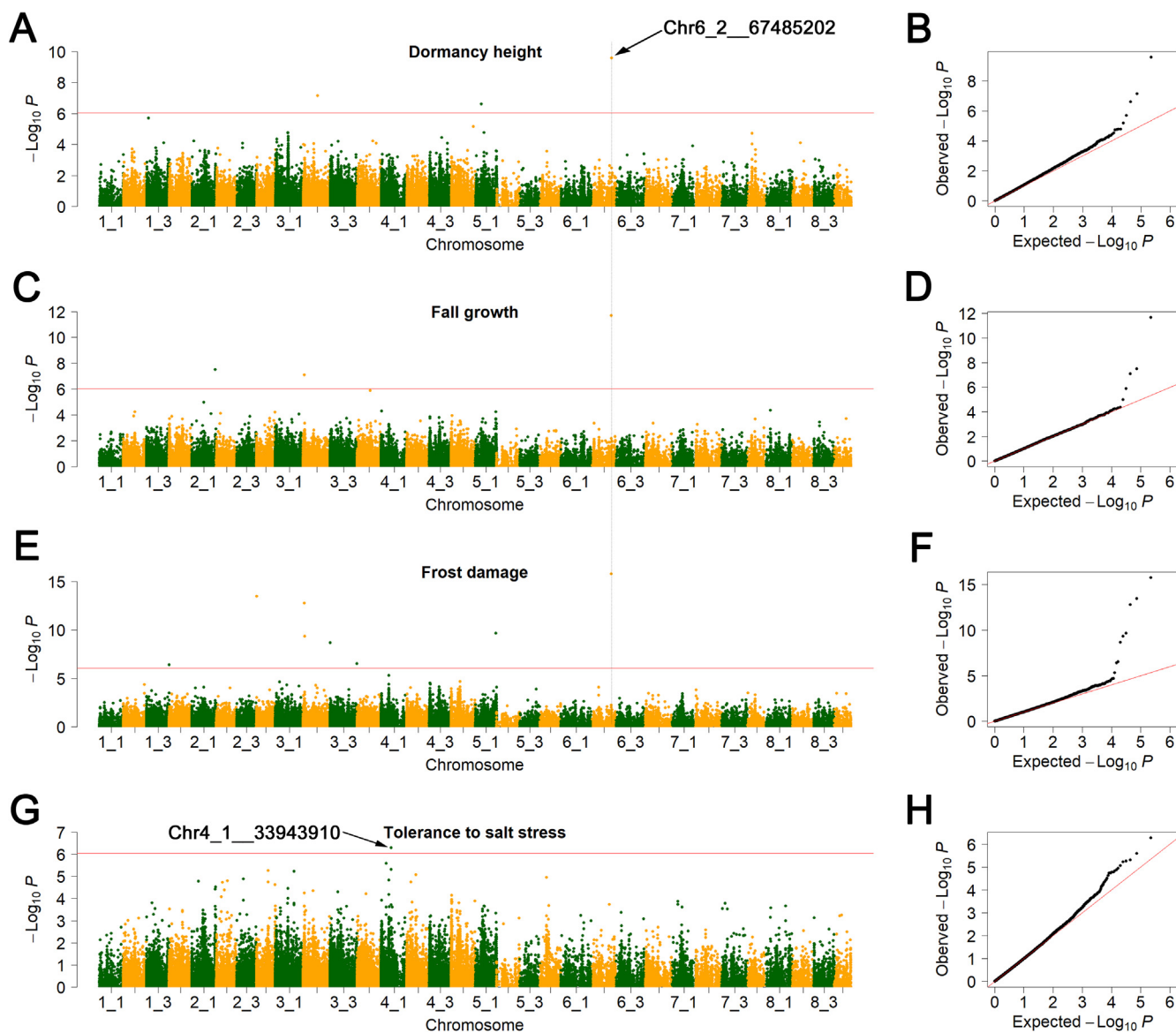

Figure 5 Manhattan plot and QQ plot of the SNPs associated with fall dormancy and salt tolerance
A. Manhattan plot of the SNPs associated with dormancy height. **B.** QQ plot of the SNPs associated with dormancy height. **C.** Manhattan plot of the SNPs associated with fall growth. **D.** QQ plot of the SNPs associated with fall growth. **E.** Manhattan plot of the SNPs associated with frost damage. **F.** QQ plot of the SNPs associated with frost damage. In (A), (C), and (E), the dashed line indicates the location of a SNP associated with all the three traits, whereas the arrow indicates its locus. **G.** Manhattan plot of the SNPs associated with salt stress tolerance. **H.** QQ plot of the SNPs associated with salt stress tolerance. The GWAS was performed using the Blink function of the GAPIT3 software. The significance threshold was set as $P = 9.00 \times 10^{-7}$.

(LD) analysis indicated that a decay was observed at 15 kb in the whole genome with $r^2 < 0.1$ (Figure 4D).

GWAS for 93 phenotypic traits

To identify candidate loci/genes related to important alfalfa agronomic traits, 93 traits (35 collected in this study and 58 collected from the USDA GRIN database) were analyzed by a GWAS using the mapping panel for the 220 accessions. A total of 101 SNPs were identified to be associated with 27 traits, including quality-related traits (*e.g.*, crude protein content and neutral detergent insoluble crude protein content), disease-related traits (*e.g.*, resistance to aphanomyces root rot and verticillium wilt), and growth-related traits (*e.g.*, dormancy height and fall growth). Our results showed that the average number of SNPs associated with a trait was 4. Notably, the resistance to blue alfalfa aphid was associated with the most SNPs ($n = 14$), while nine traits were associated with

only a single SNP (Table S8; Figure S8). Trait-associated markers are valuable resources for the marker-assisted selection during alfalfa breeding.

We also observed several SNPs individually associated with multiple traits. For example, the SNP at 20.41 Mb on chromosome 3_1 (Chr3_1__20410799) was associated with the crude protein content, neutral detergent insoluble crude protein content, non-fiber carbohydrate content, and phosphorus content (Table S8; Figure S9), which are key indexes for hay quality and nutritional value of alfalfa. We also identified a locus at 67.48 Mb on chromosome 6_2 (Chr6_2__67485202) associated with dormancy height, fall growth, and frost damage (Figure 5A–F), which are closely related to one another. An allele frequency analysis revealed that the A allele of Chr6_2__67485202, which was significantly associated with high dormancy, was present more frequently in the selected cultivated alfalfa (0.64) and landraces (0.61) relative to the wild materials (0.44) (Figure 6A). Geographically, the A allele
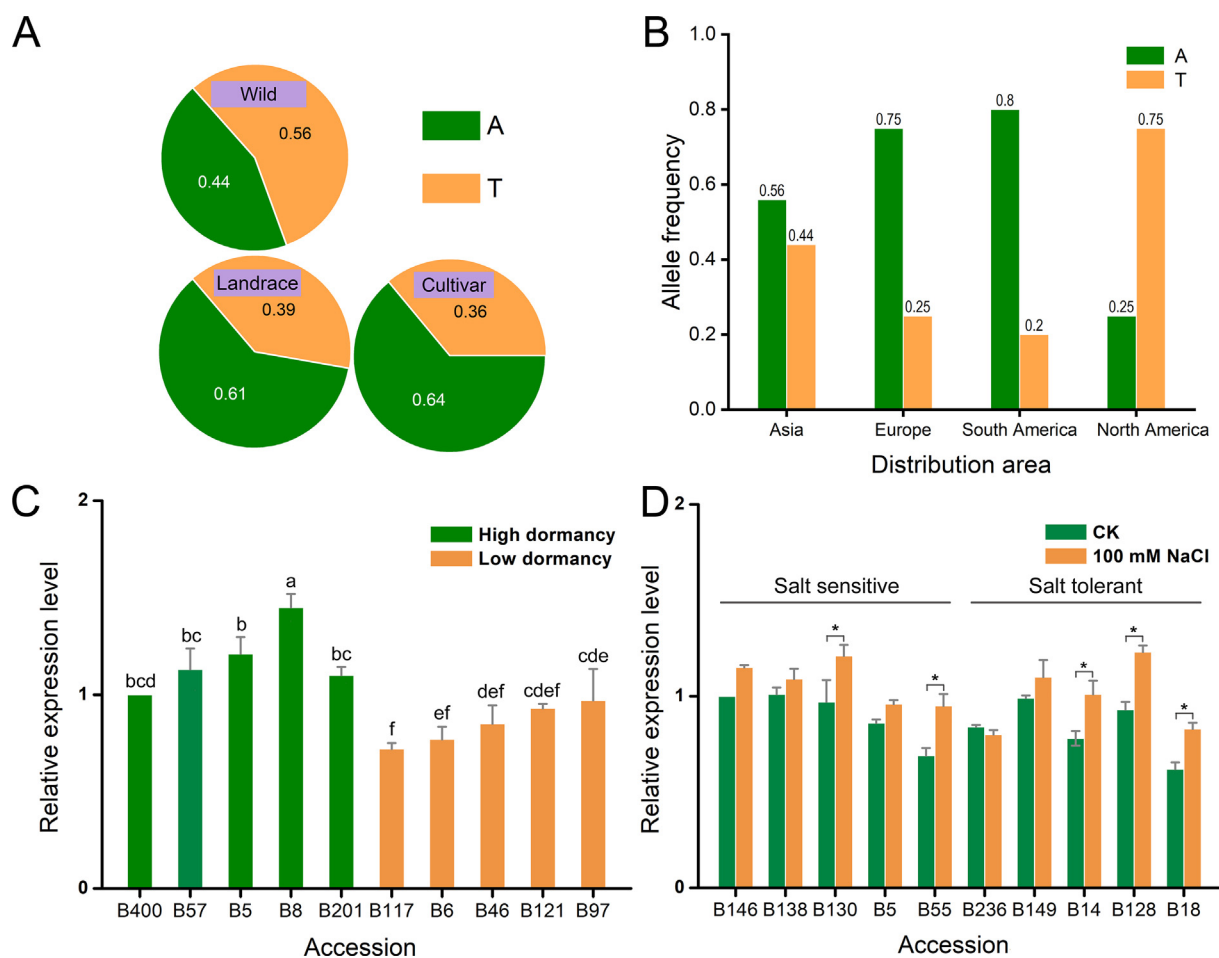

Figure 6 The allele frequency of a multi-effect SNP and the relative expression of two candidate genes in representative accessions
A. Allele frequency of a SNP (Chr6_2__67485202) in wild materials, landraces, and cultivars of alfalfa. **B.** Allele frequency of Chr6_2__67485202 in cultivars from Asia, Europe, South America, and North America. **C.** Relative expression analysis (qRT-PCR) of a candidate gene (Msa0924940) encoding a NADH dehydrogenase for SNP (Chr6_2__67485202) in representative accessions. Data were arbitrarily normalized against the data of accession B400. The different letters refer to the significant differences at $P < 0.05$ (Duncan's multiple range test) **D.** Relative expression analysis (qRT-PCR) of MYM-type zinc finger protein gene (Msa0544870) in representative salt-tolerant and salt-sensitive accessions. Data were arbitrarily normalized against the control data of accession B146. *, $P < 0.05$ (Student's *t*-test). CK, control data; NADH, nicotinamide adenine dinucleotide; MYM, myeloproliferative and mental retardation.

frequency was higher in the cultivars from Asia (0.56), Europe (0.75), or South America (0.8) than those from North America (0.25) (Figure 6B). Seven protein-coding genes were annotated within 15 kb (LD = 15 kb, r^2 = 0.1) upstream and downstream of the SNP in the Zhongmu-4. The expression data demonstrated that one candidate gene [Msa0924940, nicotinamide adenine dinucleotide (NADH) dehydrogenase] had higher expression in high-dormancy accessions than those with low-dormancy (Figure 6C). Regarding the SNP (at 33.94 Mb on chromosome 4_1) significantly associated with salt stress tolerance (Figure 5G and H), the expression of one of the two candidate genes [Msa0544870, myeloproliferative and mental retardation (MYM)-type zinc finger protein gene; Msa0544860, TPD1 protein homolog 1-like protein gene] was up-regulated by salinity treatment (100 mM NaCl), and the salt-tolerant accessions showed higher induction of Msa0544870 compared to the salt-sensitive ones (Figure 6D).

Discussion

In this study, we assembled the allele-aware chromosome-scale genome of an autotetraploid Chinese alfalfa cultivar Zhongmu-4 by using the third-generation sequencing and Hi-C sequencing technologies as well as allele-aware chromosome-level assembly algorithm. This is the second allele-aware chromosome-level reference genome assembled for an autotetraploid alfalfa species; the first was constructed for the Xinjiangdaye alfalfa landrace [28]. Shen et al. recently assembled the chromosome-level genome sequence of alfalfa cultivar Zhongmu-1, but only to the haploid chromosome level [27]. The assembly quality parameters N50, BUSCO, and anchor rate for the Zhongmu-4 assembly were better than those for the earlier allele-aware Xinjiangdaye genome assembly. Additionally, the Zhongmu-4 genome was shorter than the Xinjiangdaye genome (Table S9). These differences may be related to the diversity in the sequencing strategies used for analyzing these two species. More specifically, the PacBio CLR mode was used for the Zhongmu-4 genome sequencing, whereas the Nanopore and PacBio circular consensus sequencing (CCS) mode were used for the Xinjiangdaye genome sequencing. Nucleic acid sequence alignment of Zhongmu-4 subgenomes, Zhongmu-1 genome, and Xinjiangdaye subgenomes revealed that they had high similarity on the whole genome (Figure S10). In some regions of chromosomes, especially in chromosome 6, there were some alignment gaps, which may be caused by the high abundance of TEs in this chromosome. Because of outcrossing and self-incompatibility, alfalfa is a highly heterozygous species, with genotypic variations between alfalfa cultivars and individual plants. Accordingly, we assembled the allele-aware chromosome-level genome of Zhongmu-4, which is a representative cultivar of the Zhongmu series. Compared with the previously sequenced haploid genome of Zhongmu-1, our allele-aware genome of Zhongmu-4 has more BUSCOs and provides more information regarding sequence variations and alleles. The rapid development of sequencing technologies has enabled the construction of multiple high-quality reference genomes for some important crops, including maize and rice. These reference genomes have been used for the comprehensive identification of genome-wide variations and for functional genomics investigations and breeding practices. However, the genome resources available for alfalfa are still relatively limited. In this study, we generated a new high-quality reference genome resource for future research in alfalfa.

Alfalfa is an important legume forage crop cultivated worldwide for the subsequent use in livestock production because of its high protein content and high nutritional value [1]. Its symbiosis with *Rhizobium* bacterial species and ability to fix nitrogen are critical for protein synthesis [3]. Nitrogen is an important nutrient for plant growth and development, but it is also a crucial component of biological macromolecules, *e.g.*, nucleic acids and proteins [30]. Plants obtain nitrogen from the soil mainly via the uptake of nitrogenous compounds, *e.g.*, NO_3^- and NH_4^+, through their roots. Inorganic nitrogen must be processed into glutamine by the nitrogen metabolism pathway for further transformation and use (Figure 3C). GiS and GDH are key enzymes responsible for converting inorganic nitrogen into an organic form [31]. Interestingly, our analyses revealed that there are more *GiS* and *GDH* genes in the Zhongmu-4 genome than in the genomes of other plant species, including corn, rice, and *M. truncatula*; however, the soybean genome contains a similar number of these genes (Table S7). This is consistent with the ability of Zhongmu-4 plants to absorb nitrogen and efficiently synthesize proteins.

The lack of a reference genome was a limiting factor for earlier molecular research on alfalfa, *e.g.*, GWAS, QTL analyses, and investigations regarding evolution. Traditional methods used for breeding new alfalfa cultivars with improved traits are time-consuming processes and require the screening of large populations. A molecular marker-assisted breeding approach involving a high-quality reference genome can shorten the breeding cycle and increase breeding efficiency. To evaluate the genome quality and identify agronomic traits associated with molecular markers, we used the assembled genome of Zhongmu-4 as the reference genome for a GWAS and an analysis of the population structure of 220 accessions.

Alfalfa genetic improvements to date have mainly relied on recurrent phenotypic selection and outcrossing. The genetic exchange among individual alfalfa plants is generally unavoidable in alfalfa-growing regions [2]. Distinguishing accessions according to geographical origins is more difficult for alfalfa than for self-pollinated plants (*e.g.*, soybean and rice) [32,33]. In the PCA conducted in this study, the top three PCs only explained 6.8% of the genetic variance, indicating a weak population structure among alfalfa accessions. However, the population structure cluster information clarified the genetic segregation of accessions originating from different geographical regions. For example, the accessions from Asia and South America were separated into different genetic clusters (Figure 4B), which is consistent with the results of an earlier study on the alfalfa core germplasm population [27]. These genetic clusters may be explained by the differences in the domestication and adaptations to environmental conditions among the alfalfa accessions from diverse regions. However, outcrossing and the applied breeding strategy resulted in substantial gene flow between accessions. The genes mediating adaptations to different regions may be exchanged, thereby enabling alfalfa to spread worldwide. Furthermore, the European accessions were genetically diverse, which may be related to the fact that alfalfa originated in the Caucasus region (*i.e.*, northeastern Turkey, Turkmenistan, and northwestern Iran). This phenomenon was reported previously [6,34,35].

Because of the weak population structure, it is easier to control false positive SNPs in alfalfa than in self-pollinated crops, including rice [36] and soybean [37]; this was reflected in our QQ plot results (Figure 5, Figure S8). Thus, alfalfa research does not require the construction of an artificial population to control the population structure. However, heterozygosity is a major problem for an alfalfa GWAS. We detected SNP markers using single-allele chromosomes. Homozygous genotypes among individuals were detected for 57% of the markers. Heterozygous genotypes make it difficult to accurately estimate the MAF and filter MAF markers. Furthermore, the GWAS method depends on a homozygous genotype for SNPs among individuals. Too many heterozygotes may lead to inaccurate GWAS results for the markers [6,24,38]. Although there are some challenges to conduct an alfalfa GWAS, we still identified 101 SNPs significantly associated with 27 phenotypes. The subsequent analysis of one SNP (Chr6_2__67485202) revealed differences in the allele frequency between cultivars/landraces and wild accessions. The allele frequency also differed between cultivars from different geographical regions. These findings were consistent with the domestication and crop improvement patterns of maize [39] and soybean [33]. Furthermore, the expression level of a candidate gene (Msa0924940) located close to the SNP Chr6_2__67485202 was higher in high-dormancy accessions than in low-dormancy accessions. The Msa0924940 homologs in *M. truncatula* encode NADH dehydrogenases. Some members of this gene family were reported to mediate cold and some other environment stresses in *Arabidopsis* and potato [40,41]. Dormancy is highly correlated with low-temperature stress responses in alfalfa [42]. Accordingly, the NADH dehydrogenase encoded by Msa0924940 may be crucial for regulating the dormancy and low-temperature stress responses of alfalfa. Our gene expression analysis indicated that the expression of Msa0544870, which is located close to the salt stress-associated SNP Chr4_1__33943910 and encodes the MYM-type zinc finger protein, was induced by salt stress. However, the expression of this gene did not differ significantly between salt-tolerant and salt-sensitive accessions. The expression of a homologous zinc finger protein-coding genes in *Arachis hypogaea* was reportedly induced by salt stress, leading to enhanced salt stress tolerance [43]. Thus, Msa0544870 likely positively regulates salt stress responses in alfalfa. However, the functions of the associated genes in alfalfa need to be validated. Considered together, the SNPs identified in this study may be useful for future molecular characterization of alfalfa as well as for molecular breeding and investigations of candidate genes.

Materials and methods

Plant materials and genomic DNA sequencing

Fresh leaves, stems, and flowers were collected from one 5-year-old Zhongmu-4 alfalfa cultivar plant grown at the planting base, Langfang, Hebei Province, China (39.59 °N, 116.59 °E). Genomic DNA was extracted from the leaves using the PlantZol Kit (Catalog No. EE141-01, TransGen Biotech, Beijing, China). A portion of the DNA was used to construct 20-kb CLR libraries, which were sequenced using the PacBio Sequel platform (Pacific Biosciences,

San Francisco, CA). Nineteen single-molecule real-time cells were sequenced. The remaining DNA was used to construct libraries for the paired-end sequencing using the Illumina NovaSeq platform (Illumina, San Diego, CA). The sequencing was completed by Biomarker Technologies Corporation (Beijing, China) and Novogene Corporation (Beijing, China).

A total of 220 alfalfa accessions, including 55 cultivars, 26 cultivated materials, 95 landraces, 4 breeding materials, 16 wild materials, and 24 materials with an unclear improvement status (Table S10), were used for a GWAS. Twenty-six of the accessions were obtained from the Medium Term Library of the National Grass Seed Resources of China, whereas 194 accessions were obtained from the U.S. National Plant Germplasm System (https://npgsweb.ars-grin.gov/gringlobal/search). The accessions were mainly from China, USA, Turkey, Afghanistan, Uzbekistan, Spain, Russia, Morocco, France, and Argentina. In 2018, the 220 accessions were grown in a field plot at the Chinese Academy of Agricultural Sciences in Langfang, Hebei Province, China according to a randomized complete block design with three replications. For each accession, five plants (*i.e.*, replicates) were grown in a single row, with 30 cm between plants and 65 cm between rows. To maintain uniform plants, they were clipped to a height of 5 cm after transplanting. All plants were clipped again at the early flowering stage (*i.e.*, when 10% of the plants started to flower). A total of 93 traits were analyzed in the GWAS, including 4 agronomic traits (fall growth, dormancy, leaf length, and leaf width), 31 nutrient-related traits, and 58 traits from the U.S. National Plant Germplasm System (31 growth-related phenotypes and 27 stress resistance-related phenotypes) (Table S11).

Hi-C library construction and sequencing

The Hi-C libraries were constructed as previously described [44]. Young leaves collected from Zhongmu-4 plant were fixed in formaldehyde. The denatured chromatin was isolated from the fixed leaves and digested with *Hind* III. The ends of the digested fragments were labeled with biotin and ligated to form chimeric junctions. The chimeric fragments were isolated and then used to construct five Hi-C libraries that were sequenced using the Illumina NovaSeq platform to generate 150 bp paired-end reads. A total of 951 million reads (285 Gb) were produced.

Genome estimation and assembly

The Zhongmu-4 genome size was estimated according to the *K*-mer frequency-based method using the Illumina paired-end sequencing reads by KmerGenie (v1.7051) [45]. The genome size was also determined by flow cytometry. Fresh leaves from Zhongmu-4 and Jemalong A17 (a close relative with a completely sequenced genome comprising 425 Mb) plants were finely chopped using a sharp blade in 250 μl Cystain PI Absolute P nuclei extraction buffer (Sysmex, Kobe, Japan). After incubating for 60 s, the homogenate was filtered through a strainer (50 μm pores). The filtrate was mixed with 500 μl staining buffer (50 μg/ml propidium iodide and 5 U/ml RNase A), and then incubated for 15 min at room temperature in darkness. The fluorescence intensities of the nuclear DNA were measured using the LSRFortessa platform

(BD Biosciences, Franklin Lakes, NJ). The Zhongmu-4 genome size was determined by comparing the average propidium iodide adsorbing value with that of Jemalong A17. The genome size was also estimated using the Plant DNA C-values Database (https://cvalues.science.kew.org/).

The genome was assembled as follows: 1) PacBio CLR sub-reads were assembled into contigs using the default parameters of Canu (v1.8) [46]. The genome size was set as 3000 Mb. 2) The obtained contigs were polished using the Illumina genome data with Pilon (v1.23), interrupted by Juicer (v1.5) according to the Hi-C data, and then preliminarily assembled into scaffolds using 3D-DNA (v180419) [47,48]. 3) The obtained scaffolds were assembled into chromosomes using the chromosome-scale autopolyploid genome assembly program ALLHiC as described on GitHub [25,26] (https://github.com/-tangerzhang/ALLHiC). The gene annotation data for *M. truncatula* Mt4.0v1 in the Phytozome database (v12) were used for assembling chromosomes and analyzing collinearity [22]. The GMAP program (v2013-10-28) was used to generate an allelic contig table to remove background Hi-C signals [49]. Synteny plots were drawn using MCScanX [50].

Gene annotation

RNA-seq data and the amino acid sequences of proteins from *M. truncatula* (Mt4.0v1) were used for the *ab initio* prediction of genes. The RNA extracted from the leaves, stems, and flowers was reversely transcribed to cDNA, which was then used to construct paired-end libraries that were sequenced using the Illumina NovaSeq platform. The RNA-seq data were assembled using the default parameters of Trinity (v2.9.0) for the *de novo* and genome-guided assembly pipelines [51]. Two rounds of gene annotations were performed. More specifically, the assembled transcripts and homologous protein sequences were first annotated using BRAKER (v2.15), which enabled the fully automated training of the gene prediction tools GeneMark-EX (v4.61) and AUGUSTUS (v3.3.3) to analyze the RNA-seq data and/or the homologous protein sequences [52–54]. In the second round, the previously trained GeneMark and AUGUSTUS files, assembled transcripts, and homologous proteins were imported into the MAKER (v2.31.10) pipeline (http://www.yandell-lab.org/software/maker.html) to integrate the transcript evidence and the protein-coding evidence. The BUSCO program (v4) was used for evaluating annotation completeness [55]. The allelic genes were identified by AlleleFinder (https://github.com/sc-zhang/AlleleFinder).

Repetitive sequence prediction

A *de novo* TE sequence library was constructed for the assembled genome using RepeatModeler (v2.0.1, http://www.repeat-masker.org/RepeatModeler/), which employs three *de novo* repetitive sequence annotation programs (RECON, v1.08; RepeatScout, v1.0.6; LTR_retriever, v2.8.7) to identify repetitive elements in the genome sequence. The repetitive sequence library customized by RepeatModeler was then imported into RepeatMasker (v4.0.9, http://www.repeatmasker.org/) to identify and cluster the repetitive elements. Tandem Repeat Finder (v4.09) was used to detect tandem repeats [56]. To more precisely identify LTRs, the LTR_FINDER (v1.07) and

LTRharvest (a module of genometools v1.5.10) programs were employed to examine the LTRs [57–59]. Additionally, LTR_retriever (v2.8.7) was used to integrate the results of LTR_FINDER and LTRharvest. It was also used to calculate the LAI to assess the quality of the assembled genome [29].

Phylogeny and gene family evolution analyses

The sequences of the proteins encoded by genes in the first set of subgenomes (Chrx_1) of Zhongmu-4 and 14 other plant species (*Lotus japonicus*, *Pisum sativum*, *A. thaliana*, *A. hypogaea*, *Cicer arietinum*, *Glycine max*, *Glycine soja*, *Lupinus albus*, *M. truncatula*, *Phaseolus lunatus*, *Phaseolus vulgaris*, *Trifolium pratense*, *Vigna unguiculata*, and *P. trichocarpa*) were used to analyze phylogenetic relationships using OrthoFinder (v2.3.11) [60]. On the basis of the specie phylogenetic tree and the divergence time of *A. thaliana* and *M. sativa* (106 MYA) (http://timetree.org/), an ultrametric tree (timetree) was produced using the r8s program (v1.81). The changes in gene family size in these species were analyzed using CAFE (v4.2.1) [61]. The key enzymes in the nitrogen metabolism pathway were identified by searching the KEGG database [62]. A phylogenetic tree for the GiSs encoded by genes in the first subgenomes of Zhongmu-4 and seven other species was constructed according to the Neighbor-joining method of MEGA X [63]. The genome annotation data for *L. japonicus* were downloaded from an online database (http://www.kazusa.or.jp/lotus/). The genome annotation data of *P. sativum* were downloaded from an online database (https://urgi.versailles.inra.fr/Species/Pisum). The genome annotation data for *A. thaliana*, *A. hypogaea*, *C. arietinum*, *G. max*, *G. soja*, *L. albus*, *M. truncatula*, *P. lunatus*, *P. vulgaris*, *T. pratense*, *V. unguiculata*, and *P. trichocarpa* were downloaded from the Phytozome database (https://phytozome-next.jgi.doe.gov/).

Synteny and whole-genome duplication analyses

The syntenic blocks between the Zhongmu-4 subgenomes, as well as between the first Zhongmu-4 subgenome and the *M. truncatula* genome, were identified using MCScanX [50]. The Ka and Ks values of the syntenic paralogs were calculated using the Perl script "add_ka_and_ks_to_collinearity.pl" of MCScanX. Whole-genome alignment was performed by MUMmer4 [64].

Gene functional annotation and classification

The protein-coding genes identified in Zhongmu-4 were functionally annotated using eggNOG-mapper (v5.0) [65] and InterProScan (v5.36-75.0) [66]. The KEGG pathways associated with the annotated genes were identified using the KEGG automatic annotation server [67].

Resequencing and SNP calling

For each accession, young leaves (2 weeks after regrowth) were collected from a representative plant for resequencing. Total DNA was extracted from the leaves using the CWBIO Plant

Genomic DNA Kit (Catalog No. CW0531M, CoWin Biosciences, Beijing, China). At least 6 μg genomic DNA from each accession was used to construct a sequencing library according to the manufacturer's instructions (Illumina, San Diego, CA). Paired-end sequencing libraries were sequenced using the Illumina NovaSeq 6000 platform (Illumina, San Diego, CA) by Berry Genomics to generate 150-bp paired-end reads. Approximately 10 Gb sequencing data, with an average Q30 of 85%, were produced for each accession. The paired-end sequencing reads were mapped to the assembled Zhongmu-4 genome using the default parameters of BWA-MEM [68]. SAMtools was used to convert the SAM file to a BAM file and for sorting the BAM file according to the default parameters [69]. Picard was used to mark duplicate reads, whereas GenomeAnalysisTK was used to correct indels, which can be mistaken for SNPs. Additionally, SAMtools mpileup and VarScan were used to detect SNPs [70]. Furthermore, the following criteria were used to filter the SNP data using Vcftools [71]: missing rate $\leq$ 10%, MAF $\geq$ 0.05, and mean read depth $\geq$ 5.

Population structure and LD analyses

The population structure was analyzed using the default parameters of Admixture [72]. The best K value was determined according to the CV error. A PCA was conducted using GAPIT3 [73]. The PCA results were combined with the information regarding the geographic origins of the accessions using the R package ggplot2 [74]. The LD was calculated using the default parameters of PopLDdecay [75]. The VCF file containing information regarding all 111,075 SNP markers was imported into PopLDdecay. The LD results for all accessions were used to estimate the LD in the alfalfa genome. The distance against the mean r^2 within 50 kb of the LD was visualized using the R package ggplot2.

Association mapping

A GWAS was conducted using the Blink function of GAPIT3 [73,76]. The first three PC values were used as fixed effects in the model to correct false positives. The threshold used for detecting significant SNPs was the Bonferroni multiple test threshold ($P = 9.00 \times 10^{-7}$). The Blink method uses iterations to select a set of markers associated with a trait. These associated markers are fitted as covariates for testing the other markers. This method is better for controlling false positives than the kinship approach. The real and simulated data indicated that the statistical power of the Blink method is greater than that of other methods (*e.g.*, MLM and FarmCPU) [76]. Manhattan plots and QQ plots of the GWAS results were produced using the R package qqman [77].

Gene expression verification by qRT-PCR

Genes were selected from the upstream and downstream candidate regions of the LD decay distance of significant SNPs. Alfalfa accessions with the five highest or lowest values for the evaluation of salt tolerance, fall growth, and frost damage were selected for analyzing gene expression. Seeds were imbibed at 4 °C for 2 days and then incubated under long-day conditions [16-h light (24 °C)/8-h dark (20 °C)] for 4 days. Then, the seedlings were transferred to half-strength Hoagland solution and grown for 10 days. To simulate salt stress, 14-day-old seedlings were treated with 150 mM NaCl under long-day conditions. The shoots were collected 3.5 days (84 h) later. To analyze fall dormancy-related gene expression, leaves were collected from five high-dormancy accessions and five low-dormancy accessions in October. Total RNA was extracted from the collected samples using the Mini BEST Plant RNA Extraction Kit (Catalog No. 9769, Takara, Dalian, China), and then cDNA was synthesized using the One-Step PrimeScript RT-PCR Kit (Catalog No. RR014A, Takara). The qRT-PCR analysis was performed using the TB Green Premix Ex Taq II Kit (Catalog No. RR820Q, Takara) and the ABI 7300 system (Applied Biosystems, Foster City, CA). Gene expression levels were calculated according to the mean threshold cycle (Ct) values of three biological replicates. The alfalfa *β-actin* gene was used as the internal reference control.

Data availability

The genome assembly and gene annotation data are available online (https://figshare.com/s/fb4ba8e0b871007a9e6c). The whole-genome assembly data have also been deposited in the Genome Warehouse [78] at the National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation (GWH: GWHBECI00000000), and are publicly accessible at https://ngdc.cncb.ac.cn/gwh/. The genome and transcriptome sequencing raw data and the population resequencing data have been deposited in the Genome Sequence Archive [79] at the National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation (GSA: CRA005190 and CRA003659, repectively), and are publicly accessible at https://ngdc.cncb.ac.cn/gsa. The genome and transcriptome sequencing raw data have also been deposited in National Center for Biotechnology Information (Project: PRJNA685277), and are publicly accessible at https://www.ncbi.nlm.nih.gov/.

CRediT author statement

Ruicai Long: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization, Project administration, Funding acquisition. **Fan Zhang:** Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization. **Zhiwu Zhang:** Methodology, Validation, Writing - review & editing. **Mingna Li:** Validation, Writing - original draft. **Lin Chen:** Methodology, Visualization. **Xue Wang:** Investigation, Visualization. **Wenwen Liu:** Investigation. **Tiejun Zhang:** Investigation. **Long-Xi Yu:** Writing - review & editing. **Fei He:** Investigation. **Xueqian Jiang:** Investigation. **Xijiang Yang:** Investigation. **Changfu Yang:** Investigation. **Zhen Wang:** Methodology, Writing - review & editing, Visualization, Project administration. **Junmei Kang:** Project administration, Funding acquisition. **Qingchuan**

Yang: Conceptualization, Resources, Supervision, Funding acquisition. All authors have read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2022.01.002.

ORCID

ORCID 0000-0001-6920-2979 (Ruicai Long)
ORCID 0000-0002-8495-7779 (Fan Zhang)
ORCID 0000-0002-5784-9684 (Zhiwu Zhang)
ORCID 0000-0002-4658-9063 (Mingna Li)
ORCID 0000-0002-7036-151X (Lin Chen)
ORCID 0000-0003-2902-1393 (Xue Wang)
ORCID 0000-0002-9902-0743 (Wenwen Liu)
ORCID 0000-0002-4503-2749 (Tiejun Zhang)
ORCID 0000-0003-3132-3155 (Long-Xi Yu)
ORCID 0000-0003-1365-5359 (Fei He)
ORCID 0000-0001-8751-2292 (Xueqian Jiang)
ORCID 0000-0002-5384-3445 (Xijiang Yang)
ORCID 0000-0001-5805-7489 (Changfu Yang)
ORCID 0000-0002-5509-0264 (Zhen Wang)
ORCID 0000-0003-1051-9028 (Junmei Kang)
ORCID 0000-0002-5926-9798 (Qingchuan Yang)

References

[1] Annicchiarico P, Barrett B, Brummer EC, Julier B, Marshall AH. Achievements and challenges in improving temperate perennial forage legumes. Crit Rev Plant Sci 2015;34:327–80.

[2] Li X, Brummer EC. Applied genetics and genomics in alfalfa breeding. Agronomy 2012;2:40–61.

[3] Redondo FJ, de la Pena TC, Lucas MM, Pueyo JJ. Alfalfa nodules elicited by a flavodoxin-overexpressing *Ensifer meliloti* strain display nitrogen-fixing activity with enhanced tolerance to salinity stress. Planta 2012;236:1687–700.

[4] National Animal Husbandry Station. China Grassland Statistics. Beijing: China Agriculture Press; 2021.

[5] Hawkins C, Yu LX. Recent progress in alfalfa (*Medicago sativa* L.) genomics and genomic selection. Crop J 2018;6:565–75.

[6] Wang Z, Wang X, Zhang H, Ma L, Zhao H, Jones CS, et al. A genome-wide association study approach to the identification of candidate genes underlying agronomic traits in alfalfa (*Medicago sativa* L.). Plant Biotechnol J 2020;18:611–3.

[7] Liu XP, Yu LX. Genome-wide association mapping of loci associated with plant growth and forage production under salt stress in alfalfa (*Medicago sativa* L.). Front Plant Sci 2017;8:853.

[8] Annicchiarico P, Nazzicari N, Li X, Wei Y, Pecetti L, Brummer EC. Accuracy of genomic selection for alfalfa biomass yield in different reference populations. BMC Genomics 2015;16:1020.

[9] Yu LX, Boge B, Hu J, Fransen S, Norberg S. Genetic basis of plasticity for forage quality traits in response to water deficit in a diverse germplasm panel of alfalfa. bioRxiv 2019;554402.

[10] Wang Z, Qiang H, Zhao H, Xu R, Zhang Z, Gao H, et al. Association mapping for fiber-related traits and digestibility in alfalfa (*Medicago sativa*). Front Plant Sci 2016;7:331.

[11] Biazzi E, Nazzicari N, Pecetti L, Brummer EC, Palmonari A, Tava A, et al. Genome-wide association mapping and genomic selection for alfalfa (*Medicago sativa*) forage quality traits. PLoS One 2017;12:e0169234.

[12] Lin S, Medina CA, Boge B, Hu J, Fransen S, Norberg S, et al. Identification of genetic loci associated with forage quality in response to water deficit in autotetraploid alfalfa (*Medicago sativa* L.). BMC Plant Biol 2020;20:303.

[13] Zhang T, Yu LX, Zheng P, Li Y, Rivera M, Main D, et al. Identification of loci associated with drought resistance traits in heterozygous autotetraploid alfalfa (*Medicago sativa* L.) using genome-wide association studies with genotyping by sequencing. PLoS One 2015;10:e0138931.

[14] Yu LX, Liu X, Boge W, Liu XP. Genome-wide association study identifies loci for salt tolerance during germination in autotetraploid alfalfa (*Medicago sativa* L.) using genotyping-by-sequencing. Front Plant Sci 2016;7:956.

[15] Yu LX. Identification of single-nucleotide polymorphic loci associated with biomass yield under water deficit in alfalfa (*Medicago sativa* L.) using genome-wide sequencing and association mapping. Front Plant Sci 2017;8:1152.

[16] McCord P, Gordon V, Saha G, Hellinga J, Vandemark G, Larsen R, et al. Detection of QTL for forage yield, lodging resistance and spring vigor traits in alfalfa (*Medicago sativa* L.). Euphytica 2014;200:269–79.

[17] Zhang F, Kang J, Long R, Yu LX, Wang Z, Zhao Z, et al. High-density linkage map construction and mapping QTL for yield and yield components in autotetraploid alfalfa using RAD-seq. BMC Plant Biol 2019;19:165.

[18] Adhikari L, Makaju SO, Missaoui AM. QTL mapping of flowering time and biomass yield in tetraploid alfalfa (*Medicago sativa* L.). BMC Plant Biol 2019;19:359.

[19] Adhikari L, Lindstrom OM, Markham J, Missaoui AM. Dissecting key adaptation traits in the polyploid perennial *Medicago sativa* using GBS-SNP mapping. Front Plant Sci 2018;9:934.

[20] Li X, Alarcón-Zúñiga B, Kang J, Nadeem Tahir MH, Jiang Q, Wei Y, et al. Mapping fall dormancy and winter injury in tetraploid alfalfa. Crop Sci 2015;55:1995–2011.

[21] Li X, Wei Y, Acharya A, Hansen JL, Crawford JL, Viands DR, et al. Genomic prediction of biomass yield in two selection cycles of a tetraploid alfalfa breeding population. Plant Genome 2015;8: eplantgenome2014.12.0090.

[22] Tang H, Krishnakumar V, Bidwell S, Rosen B, Chan A, Zhou S, et al. An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. BMC Genomics 2014;15:312.

[23] Sakiroglu M, Brummer EC. Identification of loci controlling forage yield and nutritive value in diploid alfalfa using GBS-GWAS. Theor Appl Genet 2017;130:261–8.

[24] Yu LX, Zheng P, Zhang T, Rodringuez J, Main D. Genotyping-by-sequencing-based genome-wide association studies on Verti-

cillium wilt resistance in autotetraploid alfalfa (*Medicago sativa* L.). Mol Plant Pathol 2017;18:187–94.

[25] 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 2019;5:833–45.

[26] Zhang J, Zhang X, Tang H, Zhang Q, Hua X, Ma X, et al. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. Nat Genet 2018;50:1565–73.

[27] Shen C, Du H, Chen Z, Lu H, Zhu F, Chen H, et al. The chromosome-level genome sequence of the autotetraploid alfalfa and resequencing of core germplasms provide genomic resources for alfalfa research. Mol Plant 2020;13:1250–61.

[28] Chen H, Zeng Y, Yang Y, Huang L, Tang B, Zhang H, et al. Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. Nat Commun 2020;11:2494.

[29] Ou S, Chen J, Jiang N. Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res 2018;46:e126.

[30] Bloom AJ. The increasing importance of distinguishing among plant nitrogen sources. Curr Opin Plant Biol 2015;25:10–6.

[31] Martin A, Lee J, Kichey T, Gerentes D, Zivy M, Tatout C, et al. Two cytosolic glutamine synthetase isoforms of maize are specifically involved in the control of grain production. Plant Cell 2006;18:3252–74.

[32] Huang X, Zhao Y, Wei X, Li C, Wang A, Zhao Q, et al. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat Genet 2012;44:32–9.

[33] Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol 2015;33:408–14.

[34] Sakiroglu M, Doyle JJ, Brummer EC. Inferring population structure and genetic diversity of broad range of wild diploid alfalfa (*Medicago sativa* L.) accessions using SSR markers. Theor Appl Genet 2010;121:403–15.

[35] Qiang H, Chen Z, Zhang Z, Wang X, Gao H, Wang Z. Molecular diversity and population structure of a worldwide collection of cultivated tetraploid alfalfa (*Medicago sativa* subsp sativa L.) germplasm as revealed by microsatellite markers. PLoS One 2015;10:e0124592.

[36] Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 2010;42:961–7.

[37] Fang C, Ma Y, Wu S, Liu Z, Wang Z, Yang R, et al. Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. Genome Biol 2017;18:161.

[38] Jia C, Wu X, Chen M, Wang Y, Liu X, Gong P, et al. Identification of genetic loci associated with crude protein and mineral concentrations in alfalfa (*Medicago sativa*) using association mapping. BMC Plant Biol 2017;17:97.

[39] Guo L, Wang X, Zhao M, Huang C, Li C, Li D, et al. Stepwise *cis*-regulatory changes in *ZCN8* contribute to maize flowering-time adaptation. Curr Biol 2018;28:3005–15.

[40] Orzechowski S, Sitnicka D, Grabowska A, Compart J, Fettke J, Zdunek-Zastocka E. Effect of short-term cold treatment on carbohydrate metabolism in potato leaves. Int J Mol Sci 2021;22:7203.

[41] Sweetman C, Waterman CD, Rainbird BM, Smith PMC, Jenkins CD, Day DA, et al. AtNDB2 is the main external NADH dehydrogenase in mitochondria and is important for tolerance to environmental stress. Plant Physiol 2019;181:774–88.

[42] Djaman K, Owen C, Koudahe K, O'Neill M. Evaluation of different fall dormancy-rating alfalfa cultivars for forage yield in a semiarid environment. Agronomy 2020;10:146.

[43] Sui J, Li G, Chen G, Zhao C, Kong X, Hou X, et al. RNA-seq analysis reveals the role of a small GTP-binding protein, Rab7, in

regulating clathrin-mediated endocytosis and salinity-stress resistance in peanut. Plant Biotechnol Rep 2017;11:43–52.

[44] Wang C, Liu C, Roqueiro D, Grimm D, Schwab R, Becker C, et al. Genome-wide analysis of local chromatin packing in *Arabidopsis thaliana*. Genome Res 2015;25:246–56.

[45] Chikhi R, Medvedev P. Informed and automated *k*-mer size selection for genome assembly. Bioinformatics 2014;30:31–7.

[46] Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res 2017;27:722–36.

[47] Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science 2017;356:92–5.

[48] Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst 2016;3:95–8.

[49] Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 2005;21:1859–75.

[50] Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res 2012;40:e49.

[51] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 2011;29:644–52.

[52] Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-genome annotation with BRAKER. Methods Mol Biol 2019;1962:65–95.

[53] Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics 2008;24:637–44.

[54] Bruna T, Lomsadze A, Borodovsky M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genom Bioinform 2020;2:lqaa026.

[55] Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;31:3210–2.

[56] Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 1999;27:573–80.

[57] Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 2007;35:W265–8.

[58] Gremme G, Steinbiss S, Kurtz S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Trans Comput Biol Bioinform 2013;10:645–56.

[59] Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. BMC Bioinformatics 2008;9:18.

[60] Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 2015;16:157.

[61] Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol 2013;30:1987–97.

[62] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2017;45:D353–61.

[63] Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 2018;35:1547–9.

[64] Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A, et al. MUMmer4: a fast and versatile genome alignment system. PLoS Comput Biol 2018;14:e1005944.

[65] Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 2019;47:D309–14.

[66] Zdobnov EM, Apweiler R. InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics 2001;17:847–8.

[67] Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 2007;35:W182–5.

[68] Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 2013;1303.3997.

[69] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078–9.

[70] Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 2012;22:568–76.

[71] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics 2011;27:2156–8.

[72] Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 2009;19:1655–64.

[73] Wang J, Zhang Z. GAPIT version 3: boosting power and accuracy for genomic association and prediction. Genomics Proteomics Bioinformatics 2021;19:629–40.

[74] Villanueva RAM, Chen ZJ. ggplot2: elegant graphics for data analysis. Meas Interdiscip Res Perspect 2019;17:160–7.

[75] Zhang C, Dong SS, Xu JY, He WM, Yang TL, Schwartz R. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics 2019;35:1786–8.

[76] Huang M, Liu X, Zhou Y, Summers RM, Zhang Z. BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. Gigascience 2019;8:giy154.

[77] Turner SD. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. bioRxiv 2014;005165.

[78] Chen M, Ma Y, Wu S, Zheng X, Kang H, Sang J, et al. Genome Warehouse: a public repository housing genome-scale data. Genomics Proteomics Bioinformatics 2021;19:584–9.

[79] Chen T, Chen X, Zhang S, Zhu J, Tang B, Wang A, et al. The Genome Sequence Archive Family: toward explosive data growth and diverse data types. Genomics Proteomics Bioinformatics 2021;19:578–83.