CellPress
Partner Journal

Molecular Plant
Resource Article

The Chromosome-Level Genome Sequence of the Autotetraploid Alfalfa and Resequencing of Core Germplasms Provide Genomic Resources for Alfalfa Research

Chen Shen[1,4], Huilong Du[2,3,4], Zhuo Chen[2,3,4], Hongwei Lu[2,3,4], Fugui Zhu[1], Hong Chen[1], Xiangzhao Meng[1], Qianwen Liu[1], Peng Liu[1], Lihua Zheng[1], Xiuxiu Li[2,3], Jiangli Dong[1,*], Chengzhi Liang[2,3,*] and Tao Wang[1,*]

[1]State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing, China

[2]State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Innovation Academy for Seed Design, Chinese Academy of Sciences, Beijing, China

[3]University of Chinese Academy of Sciences, Beijing, China

[4]These authors contributed equally to this article.

***Correspondence: Jiangli Dong (dongjl@cau.edu.cn), Chengzhi Liang (cliang@genetics.ac.cn), Tao Wang (wangt@cau.edu.cn)**

https://doi.org/10.1016/j.molp.2020.07.003

ABSTRACT

Alfalfa (*Medicago sativa*) is one of the most important forage crops in the world; however, its molecular genetics and breeding research are hindered due to the lack of a high-quality reference genome. Here, we report a *de novo* assembled 816-Mb high-quality, chromosome-level haploid genome sequence for 'Zhongmu No.1' alfalfa, a heterozygous autotetraploid. The contig N50 is 3.92 Mb, and 49 165 genes are annotated in the genome. The alfalfa genome is estimated to have diverged from *M. truncatula* approximately 8 million years ago. Genomic population analysis of 162 alfalfa accessions revealed high genetic diversity, weak population structure, and extensive gene flow from wild to cultivated alfalfa. Genome-wide association studies identified many candidate genes associated with important agronomic traits. Furthermore, we showed that *MsFTa2*, a *Flowering Locus T* homolog, whose expression is upregulated in salt-resistant germplasms, may be associated with fall dormancy and salt resistance. Taken together, these genomic resources will facilitate alfalfa genetic research and agronomic improvement.

Key words: alfalfa, genome assembly, population genetics, GWAS, *Flowering Locus T*

Shen C., Du H., Chen Z., Lu H., Zhu F., Chen H., Meng X., Liu Q., Liu P., Zheng L., Li X., Dong J., Liang C., and Wang T. (2020). 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.

INTRODUCTION

Alfalfa (*Medicago sativa*), known as the "Queen of Forage," is an essential perennial legume that provides inexpensive, nutritious, and highly digestible forage; the crude protein content of alfalfa hay can exceed 20% (Russelle, 2014; Butler, 1995; Elfaki and Abdelatti, 2018). Alfalfa is self-incompatible and can be crossed with different subspecies; its rich genetic diversity permits it to be widely grown under various environmental conditions (Michaud et al., 1988). Therefore, alfalfa is not only a high-quality feed for livestock, but also an important crop for environmental protection (Radović et al., 2009). As the primary animal feed for dairy cows, beef cattle, and other farm animals, alfalfa is an important basis for the prosperity of the dairy industry. The liquid milk consumption per capita in the United States (66 kg) was seven times more than that in China in 2019 (https://www.clal.it/en/?section=tabs_consumi_procapite), and the alfalfa industry therefore presents an important limitation to the development of the dairy industry in China. Because it has access to symbiotically fixed N_2, alfalfa requires reduced nitrogen fertilizer inputs, thereby producing economic and ecological benefits. Despite these advantages, drawbacks associated with alfalfa include bloating among ruminant animals, susceptibility to root disease, and sensitivity to soil pH. To improve global food

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and IPPE, CAS.

supplies and food security, it is necessary to facilitate the development of the alfalfa industry through breeding efforts.

The *M. sativa* complex is composed of perennial, outcrossing, morphologically differentiated, but often interfertile taxa (Havananda, 2010). The *M. sativa* complex includes diploid and autotetraploid subspecies. For instance, *M. sativa* ssp. *caerulea* is a diploid subspecies that has been identified as the ancestor of tetraploid cultivated alfalfa (Yu et al., 2017). By comparison, cultivated alfalfa (*M. sativa* ssp. *sativa*, 2n = 4× = 32) is a true autotetraploid that shows quadrivalents during meiosis and tetrasomic inheritance (Cao et al., 2004). These subspecies hybrids display heterosis for many quantitative traits (Bhandari et al., 2007; Li et al., 2009a, 2009b).

For years, alfalfa breeding efforts were hampered by a lack of information about its genome structure and the genetic basis of important traits (Annicchiarico et al., 2016). The study of alfalfa genomics is slowed down by its complexity in comparison with that of other related species, such as *Lotus japonicas* (Sato et al., 2008), *Glycine max* (Schmutz et al., 2010), and *Medicago truncatula* (Pecrix et al., 2018). Although quantitative trait loci and association mapping studies have been performed with simple sequence repeat (SSR) and single-nucleotide polymorphism (SNP) markers (Yu et al., 2016; Jia et al., 2017; Hawkins and Yu, 2018), few genes have been characterized (Barros et al., 2019), and the impacts of improved cultivars have been limited. Therefore, resolving the structure of the alfalfa genome is important for supporting alfalfa genetic and genomic research and for accelerating genomic selection breeding efforts (Hawkins and Yu, 2018). The main difficulty lies in the acquisition of long contigs from autopolyploid or highly heterozygous plants, such as sugarcane and strawberry, because it is difficult to simultaneously work with long repeats and heterozygosity (Zhang et al., 2018; Edger et al., 2019).

Here, we present a high-quality, chromosome-level haploid genome assembly for *M. sativa* ssp. *sativa* cv. Zhongmu No. 1, a widely grown cultivar in Northern China. This assembly consists of eight pseudo-chromosomes. We also sequenced 137 global core cultivated alfalfa germplasms and 25 ssp*. caerulea* accessions using short reads to characterize population migration history and genetic exchange between subpopulations. We identified dozens of regions associated with important agronomic traits using genome-wide association studies (GWAS). We found that the florigen gene *MsFTa2* is associated with fall dormancy, cold resistance, salt resistance, and unifoliate internode length. It is also putatively associated with alfalfa's environmental adaptability and widespread geographic distribution. Together, these resources provide a foundation for accelerating the genetic improvement of alfalfa, thereby improving global food security for a growing world population.

RESULTS

Genome Sequencing and Assembly

We sequenced the genome of the cultivated alfalfa accession Zhongmu No. 1 using the Illumina platform and Pacific Biosciences (PacBio) single-molecule real-time (SMRT) sequencing to generate approximately 528 Gb short reads and 245 Gb long reads, respectively (Supplemental Tables 1 and 2). We also used the optical map from Bionano Genomics (Lam et al., 2012) and high-throughput chromatin conformation capture (Hi-C) (Dekker, 2006) to generate 522 Gb Bionano and 299 Gb Hi-C data, respectively. K-mer analysis of the short reads showed that the alfalfa genome is highly heterozygous (Supplemental Figure 1). The estimated haploid genome size (~800 Mb) was confirmed by our 3.0–3.3 Gb 2C flow cytometry measurement (F Blondon et al., 1994; Fyad-Lameche et al., 2015) (Supplemental Figure 2). We used Canu (Koren et al., 2017) and MECAT (Xiao et al., 2017) to assemble the PacBio long reads into preliminary contigs, which were further improved to obtain extended contigs using HERA (Du and Liang, 2019) with Bionano data (Supplemental Table 3). The assembled sequences were filtered using Purge Haplotigs (Roach et al., 2018) and Redundans (Pryszcz and Gabaldón, 2016) to obtain a non-redundant genome of 816 Mb, which was close to the estimated haploid genome size. The contigs were then clustered into eight pseudo-chromosomes with the Hi-C heatmap (Supplemental Figure 3, Supplemental Tables 4 and 5). The haploid genome had a contig N50 (the minimum contig length needed to cover 50% of the genome) of 3.92 Mb (Supplemental Table 6). The non-redundancy of the genome was confirmed with the Hi-C data (Supplemental Figure 4).

We identified 1344 (93.3%) of the 1440 conserved genes in the genome assembly with BUSCO (Waterhouse et al., 2018) (Supplemental Table 7). The annotation of long terminal repeats (LTRs) revealed an LTR Assembly Index (LAI) (Ou et al., 2018) score of 22.30 that met the gold standard for high-quality reference genome. Furthermore, we found that 95.35% of the Illumina short reads mapped to the Zhongmu No. 1 genome. Taken together, these results suggest that the Zhongmu No. 1 genome is a high-quality assembly.

Genome Annotation

By combining *ab initio* prediction and transcript and protein evidence (Supplemental Tables 8 and 9), we annotated 49 165 high-confidence genes (Figure 1, Supplemental Tables 10 and 11). We annotated 76% of the genes with functional assignments (Supplemental Table 12), and approximately 92.3% of the genes were supported by Iso-Seq isoforms or the RNA sequencing (RNA-seq) data (Supplemental Table 10). As expected, gene density was higher on the arms than in the middle of the chromosomes (centromere regions) (Figure 1).

We used *ab initio* and evidence-based approaches to annotate repetitive sequences, which accounted for 57% of the genome. Similar distributions of repetitive sequences were found on all chromosomes, with a higher density in the middle of each chromosome than in the distal regions (Figure 1). The most abundant transposable elements (TEs) were LTR retrotransposons (~42%), which consisted mainly of Copia elements (~12%) and Gypsy elements (~18%) (Supplemental Table 13).

Comparative Genomic and Evolutionary Analysis

Comparative analysis of the Zhongmu No. 1 genome and the *M. truncatula* genome (Pecrix et al., 2018) showed that they were highly collinear, confirming the accuracy of our assembly (Figure 2A). There was a large translocation (~20 Mb) between chromosomes 4 and 8 and an inversion in chromosome 1 (Figure 2A and 2B). These structural variations were confirmed

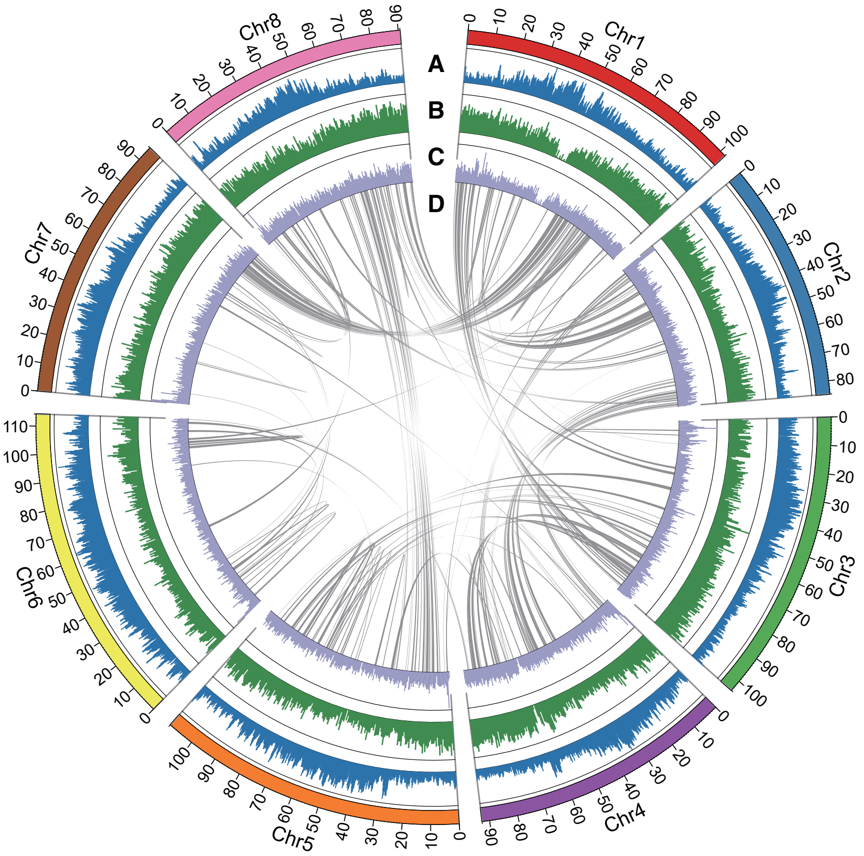

Figure 1. Distribution of Genomic Features within the Alfalfa Zhongmu No. 1 Genome.
(A–C) Circular representation of the GC content **(A)**, gene density **(B)**, and LTR density **(C)** of genome regions (100 kb for each window).
(D) Collinear gene blocks in the Zhongmu No. 1 genome.

by Zhongmu No. 1 Hi-C data (Supplemental Figure 4) and are supported by previous reports (Li et al., 2014; Pecrix et al., 2018).

We used 290 single-copy genes to build a phylogenetic tree for 10 species, including *M. sativa* and *M. truncatula*, and it showed that alfalfa is most closely related to *M. truncatula*. The divergence between *M. sativa* and *M. truncatula* was estimated to have occurred 8 million years ago (Mya) (Figure 2C). Furthermore, we used K_s values for 17 371 high-confidence (1:1 ratio only) *M. sativa* and *M. truncatula* collinear gene pairs to estimate that these species diverged from their most recent common ancestor approximately 6.4 Mya (Figure 2D).

The Zhongmu No. 1 genome is much larger than the *M. truncatula* genome. However, the collinear depth with *M. truncatula* shows a 1:1 pattern (Figure 1, Supplemental Figure 5), and Zhongmu No. 1 has not experienced obvious recent whole-genome duplication events, such as those observed in *G. max* (Supplemental Figure 6). The length of LTR-TEs in the Zhongmu No. 1 genome (~315 Mb) is much larger than that in *M. truncatula* (~65 Mb). In addition, LTR bursts occurred in Zhongmu No. 1 recently (0–0.5 Mya) after the divergence of the two species (Supplemental Figure 7). The amplification of LTR-TEs caused by LTR bursts is the major contributing factor to the alfalfa genome expansion.

Population Genomic Analysis

To understand the genetic diversity of the alfalfa population, we sequenced 137 alfalfa global core germplasms (Basigalup et al., 1995) and 25 ssp. *caerulea* accessions provided by the U.S. National Plant Germplasm System of the United States Department of Agriculture Agricultural Research Service (https://www.ars-grin.gov/npgs/index.html). The core germplasms represent a rich source of more than 50 desirable agronomic traits (Supplemental Data 1) and were acquired from specimens distributed over six continents (Figure 3A).

Because alfalfa includes both diploid subspecies and tetraploid subspecies, for simplicity we treated all accessions as diploids for the population analysis. Please note that this simplification will create limitations for some analyses (see Discussion for details). We combined various methods to analyze the sequenced populations (Supplemental Figure 8). We used the Zhongmu No. 1 genome as a reference to identify 963 654 biallelic SNP markers in coding sequence (CDS) regions whose missing rate is less than 5%. . We first performed population structure inference on the 162 alfalfa accessions with ADMIXTURE software, which divided the population into three groups. Group 1 was dominated by ssp. *caerulea*; group 2 contained two unknown diploid samples and some admixed tetraploid samples; and group 3 mainly contained tetraploid alfalfa. Group 3 could be further divided into three subgroups that were highly correlated with subspecies geographic distribution (Figure 3B and 3C). Principal component analysis (PCA) produced similar groups (Figure 3D–3F). However, the variation explained by the top three principal components of both analyses was fairly low (less than 10%), indicating a weak population structure and complex genetic composition in the alfalfa population.

Population split and migration analysis performed with TreeMix further implied highly weighted gene flows to the tetraploid alfalfa population from distant populations represented by the diploid outgroup, as well as unknown ancestors of tetraploid alfalfa and ssp. *caerulea* (Supplemental Figure 9, Supplemental Note 1). This may also explain the origin of significant heterozygous variation and weak population structure within tetraploid alfalfa.

Based on the average values of nucleotide polymorphism (π) and SNP differences, the genetic diversity of the group 3 population (tetraploid alfalfa) was higher than that of the group 1 population (ssp. *caerulea*) (Supplemental Table 14). The population differentiation analysis (F_{ST}) showed that the differentiation

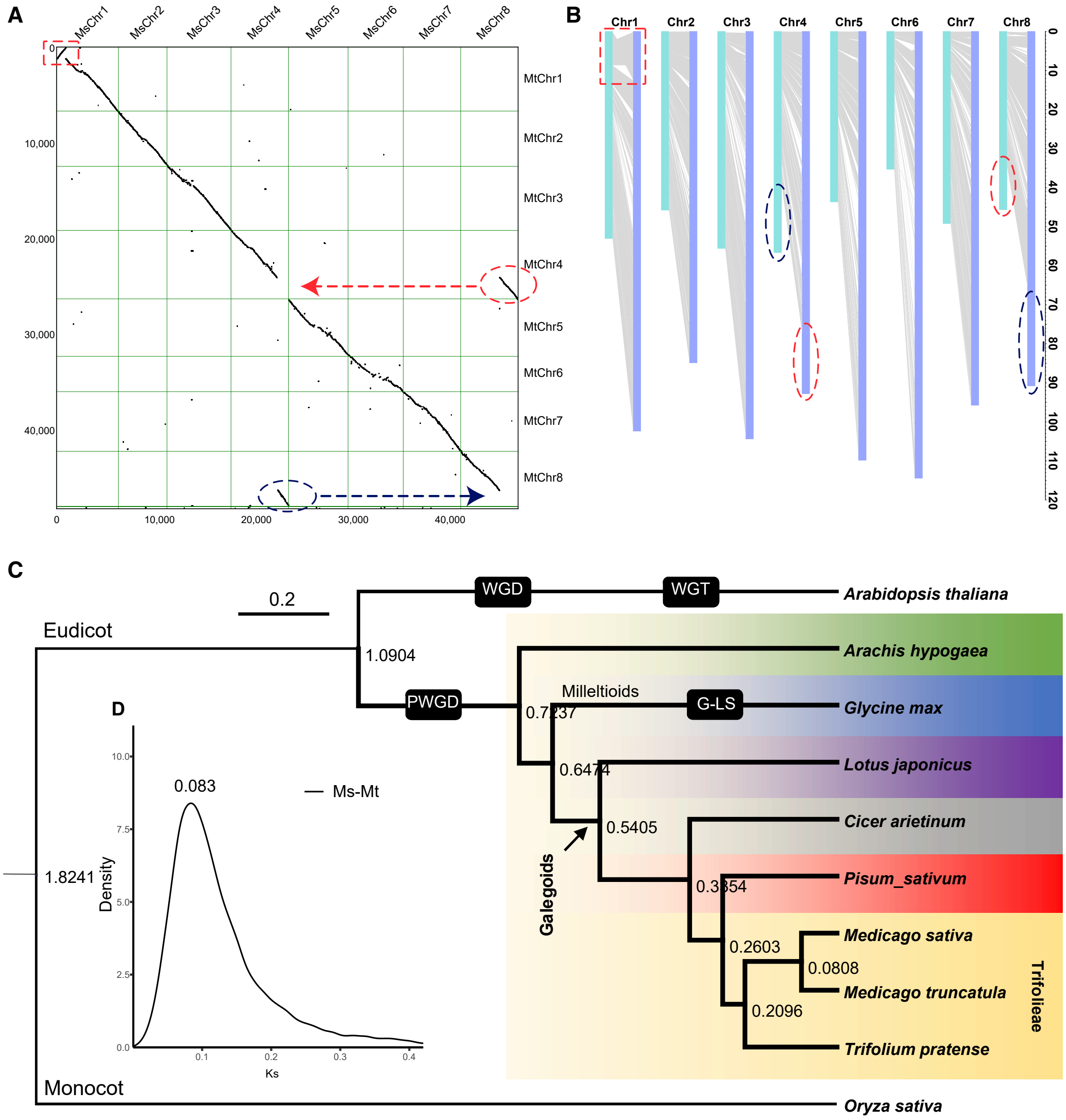

Figure 2. Alignment of *M. sativa* Zhongmu No. 1 Chromosomes with *M. truncatula* Chromosomes and Evolutionary Analysis.

(A) *M. sativa* chromosomes aligned to *M. truncatula* chromosomes. The dots indicate gene pairs.

(B) The "Zhongmu No.1" genome aligned to the *M. truncatula* genome. Green bars indicate *M. truncatula* chromosomes, and blue bars indicate *M. sativa* chromosomes. Dotted boxes depict the inversion on Chr1. Dotted circles indicate the translocation between Chr4 and Chr8.

(C) Phylogenetic tree of several plants. Branch length represents the estimated nucleotide substitutions per site. Scale bar corresponds to 0.2 substitutions per site. WGD, WGT, PWGD, and G-LS data are reproduced from the literature (Kreplak et al., 2019).

(D) Distribution of pairwise K_s for syntenic genes between *M. sativa* (Ms) and *M. truncatula* (Mt).

degree between the three tetraploid alfalfa subpopulations was high (Supplemental Table 15). Tajima's D results showed that selective sweep signals were rarely detected in the tetraploid alfalfa population (Supplemental Table 16). These results confirm that the current tetraploid alfalfa population maintains high genetic diversity and may not experience purifying selection.

North and South American alfalfa cultivars were clustered into separate groups (Figure 3A–3C), in agreement with the migration and cultivation history of alfalfa, which was brought to South America in the 16th century and was later cultivated in North America (Tysdal et al., 1942; Hanson et al., 1988) (Figure 3A). Combining historical records with our results, we suggest that cultivated alfalfa may have originated from one ancestral population in Europe. North American and South American populations were brought by two groups of colonists from Europe. Furthermore, its gene pool was continuously enriched by genetic components absorbed from other subspecies as a result of dispersion and adaptation to different environments (Supplemental Figure 9).

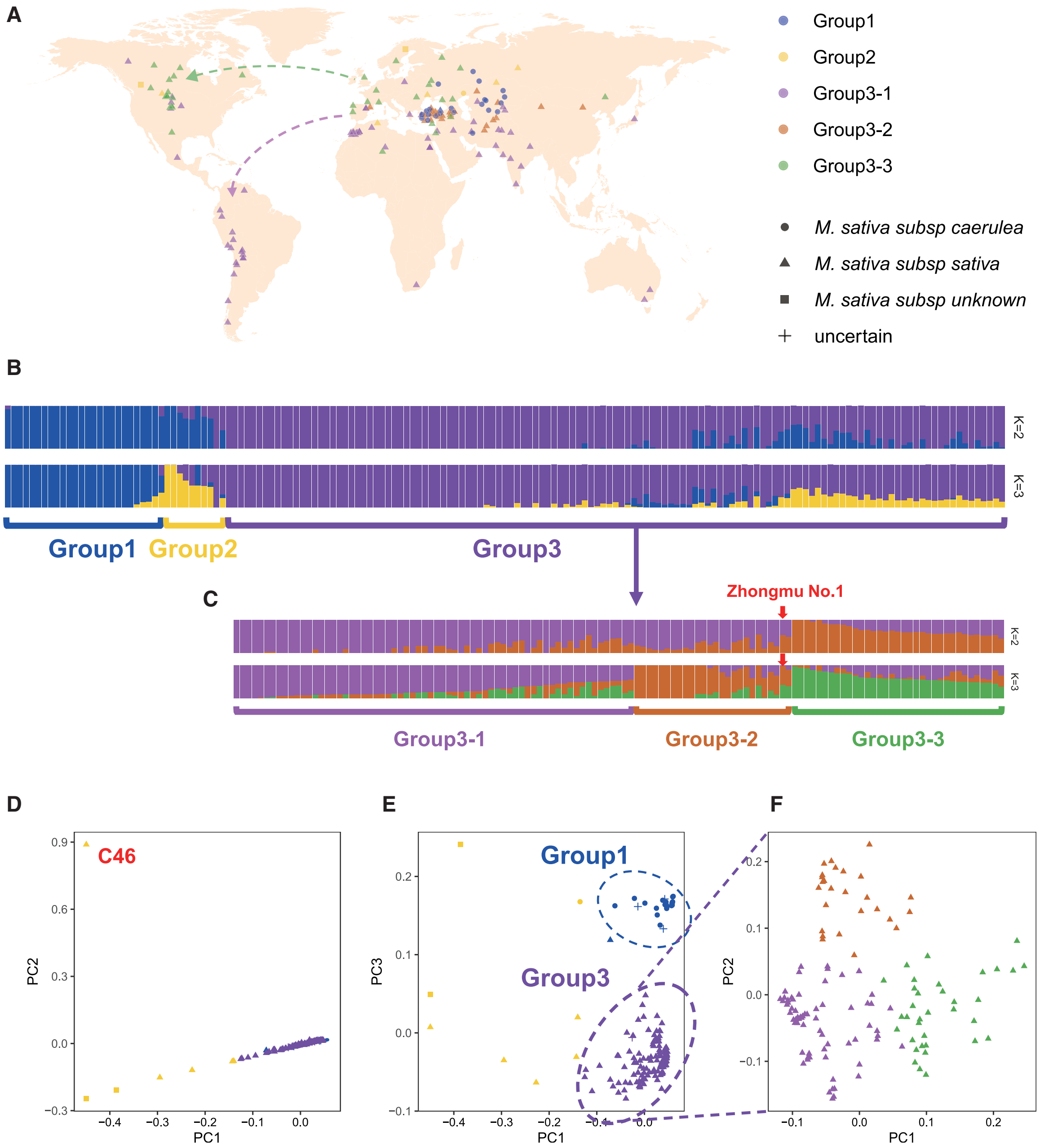

Figure 3. Geographical Distribution and Population Genetic Structure of 162 *M. sativa* Accessions.
(A) The geographical distribution of *M. sativa* on the basis of ADMIXTURE groups. Arrow indicates that a European colonist took alfalfa to America.
(B) ADMIXTURE plot of 162 *M. sativa* accessions shows three subpopulations (k = 3).
(C) ADMIXTURE plot of group 3 accessions from **(B)** shows the three subpopulations (k = 3) of group 3.
(D) Principal components (PCs) of the variation among all 162 *M. sativa* accessions with PC1 (3.8%) and PC2 (3.4%). **(E)**Principal components (PCs) of the variation among all 162 *M. sativa* accessions with PC1 (3.8%) and PC3 (2.3%).
(F) PCs of variation among group 3 accessions. PC1, 3.4%; PC2, 2.1%.

GWAS of Important Agronomic Traits

To identify candidate genes associated with key agronomic traits, we performed GWAS on the 137 global core germplasms using whole-genome SNP data and phenotypic data of the U.S. National Plant Germplasm System (Supplemental Data 1). Using 2 463 637 biallelic SNPs based on the Zhongmu No. 1 genome, we identified more than 100 candidate regions associated with more than 30 agronomic traits, including disease resistance, insect resistance, growth, morphology, productivity, and stress response (Supplemental Figure 10, Supplemental Data 3). These results identified valuable genomic markers and candidate genes that can be used in future breeding and basic research efforts.

We found a particularly interesting association locus at 73.4 Mb on chromosome 7 that was associated with four traits: fall dormancy, cold resistance, salt germination, and unifoliate

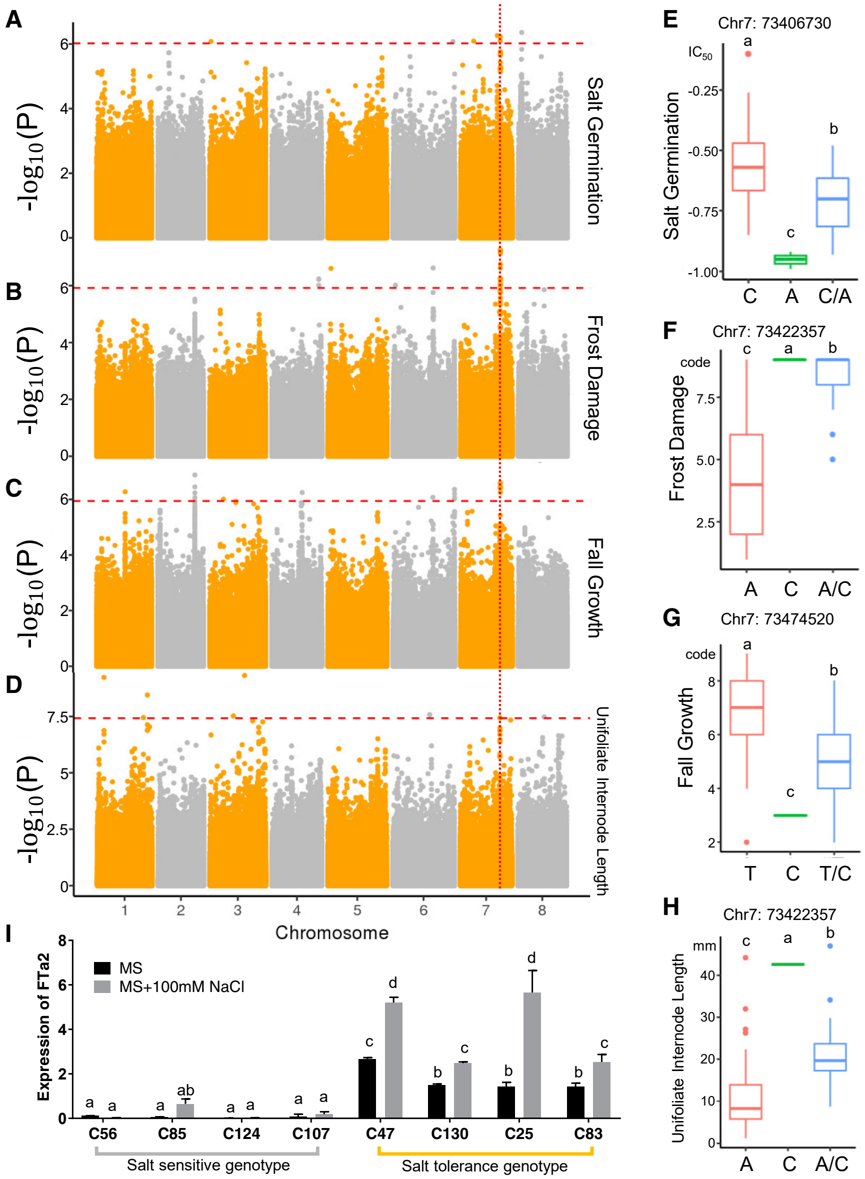

Figure 4. *FTa2* GWAS-Based Association with Four Phenotypes and Expression Analysis.

(A–D) Manhattan plot for the genome-wide association analysis of salt germination **(A)**, frost damage **(B)**, fall dormancy **(C)**, and unifoliate internode length **(D)**. The –$\log_{10}$ (*P*) values for the association tests (two-tailed) are shown on the y axis, and the chromosomes are ordered on the x axis.

(E) The SNP genotype is associated with the salt germination phenotype. Half maximal inhibitory concentration is the sodium chloride concentration that inhibits the germination of 50% of the viable seeds.

(F) The SNP genotype is associated with the frost damage phenotype. The number 1 indicates the least frost damage and 9 describes the greatest frost damage.

(G) The SNP genotype is associated with the fall dormancy phenotype. Numbers indicate fall dormancy, with larger numbers indicating slower growth in autumn and winter.

(H) The SNP genotype is associated with the unifoliate internode length phenotype. This phenotype is the distance between the cotyledonary node and the unifoliate leaf node measured in millimeters. Significant differences of **(E–H)**, indicated by letters, were determined by one-way ANOVA, $\alpha = 0.05$.

(I) The bar plot shows relative FTa2 expression in different *M. sativa* accessions with or without 100 mM NaCl treatment. C56, C85, C124, and C107 accessions have a salt-sensitive genotype. C47, C130, C25, and C83 have a salt-tolerant genotype. Error bar, SD; significant differences, indicated by letters, were determined by two-way ANOVA, $\alpha = 0.05$.

internode length (Figure 4A–4D). Stanton-Geddes et al. (2013) reported that this region was also associated with the flowering date of *M. truncatula*. The genotypes of three different candidate SNPs were highly positively correlated (*rho* = 0.85–0.98) with these four phenotypes, indicating a potential common causal mutation for all four phenotypes (Figure 4E–4H). This hypothesis is consistent with previous reports that fall dormancy is positively correlated with salt tolerance and unifoliate internode length (Liu et al., 2015, 2019a, 2019b). We found that the three SNPs were located in a gene cluster composed of three phosphatidylethanolamine binding protein (PEBP) genes (Figure 4A–4D) that are homologs of the *Arabidopsis thaliana* florigen gene *Flowering Locus T* (*FT*), which promotes the transition from vegetative growth to flowering (Kardailsky et al., 1999; Tamaki et al., 2007). A syntenic region in *M. truncatula* also contains three PEBP genes named *FTa1*, *FTa2*, and *FTc* (Supplemental Figure 11). Therefore, we named the three alfalfa PEBPs *MsFTa1*, *MsFTa2*, and *MsFTc* on the basis of sequence similarity (Supplemental Figure 12).

An *FT* homolog has been reported to be associated with tree dormancy (Cooke et al., 2012). Another *FT* homolog is induced by salt treatment, and the overexpression of this gene improved *N. tabacum* salt tolerance (Li et al., 2015). To test for salt-induced changes in *MsFT* expression, we randomly selected four salt-tolerant cultivars and four salt-sensitive cultivars for qPCR analysis. Compared with salt-sensitive accessions (C56, C85, C124, and C107) with major allele SNPs, salt-tolerant accessions (C47, C130, C25, and C83) that were heterozygous for the SNPs of interest had relatively high *FTa2* expression that was further induced by salt treatment (Figure 4I). *FTa1* and *FTc* expression patterns were not correlated with the salt sensitivity phenotypes of these specimens (Supplemental Figure 13). We found that the sequences and copy numbers of each *FT* gene did not differ among these accessions (Supplemental Figures 14 and 15), suggesting that differential *FT* expression was not caused by sequence or copy-number variation. The results of GWAS localization and expression pattern analysis suggest that *FTa2* expression differences may affect alfalfa fall dormancy and salt tolerance.

The major genotypes of these three SNPs of interest were found mainly in cultivars distributed in higher-latitude areas with low fall and winter temperatures, whereas minor genotypes were found mainly in cultivars from lower-latitude regions with higher temperatures and higher soil salinity (Supplemental Figure 16). These observations are consistent with the logic that fall dormancy is important for reducing alfalfa frost damage and high salt tolerance facilitates adaptation to lower-latitude regions. Thus, *FTa2* may contribute to the environmental adaptability of alfalfa.

DISCUSSION

We have assembled a high-quality haploid genome for autotetraploid alfalfa, the most important forage crop in the world. In the past, homozygous inbred lines were usually selected for genome assembly to reduce the influence of heterozygous sequences. Because it is very difficult to obtain homozygous lines for self-incompatible plants, we selected the major heterozygous cultivar in China, Zhongmu No. 1, as the material for *de novo sequencing*.

A highly continuous and complete set of reference genomes is an essential basis for a wide range of population genetics studies and experimental research. By combining current state-of-the-art technologies (including PacBio long read sequencing, Bionano mapping, Hi-C sequencing, and high-quality genome assemblers), we generated a representative haploid reference genome for alfalfa. The continuity and integrity of the assembled Zhongmu No. 1 genome showed its high quality: it had a contig N50 of 3.92 Mb, a BUSCO estimate of 93.3%, and an LAI score estimate of 22.30. The combined sequence length of all 8 chromosomes was 794.3 Mb, close to its estimated genome size of ~800 Mb. Because of the heterozygosity and outcrossing of alfalfa, there is undoubtedly some chimerism in the Zhongmu No. 1 reference genome. It is notable that the reference genome is not a real set of subgenomes *in vivo* and cannot be used to study the subgenome. However, its collinearity with the *M. truncatula* genome is good. Different alfalfa individuals each have their own recombination history, such that each individual is a chimera. Taking the chimera sequence as the reference will not affect the linkage relationship of genetic markers. On the premise that we cannot obtain an ideal complete non-chimeric genome at present, integrity and continuity are the most important factors that affect the quality and usability of the reference genome. We used the Zhongmu No. 1 genome as the reference genome for population structure and GWAS analysis of 162 accessions as an example.

An interesting, unique aspect of alfalfa is its multiple diploid and tetraploid subspecies. The tetraploid subspecies represented by Zhongmu No. 1 are thought to have descended from ssp. *caerulea*. Our population genomic analysis indicated that frequent introgression of genes from diploid populations reduced the population structure and increased the genetic diversity of cultivated alfalfa. We also found that alfalfa's genetic diversity was higher than that of its probable wild diploid ancestor ssp. *caerulea*. This result may have arisen because we did not sequence a sufficient number of representative ssp. *caerulea* accessions. Another possibility is that other subspecies, such as the distant ssp. *falcata*, have also contributed to alfalfa's genetic diversity.

For the population analysis, we treated SNPs using a simplified approach in which SNPs were called in a diploid manner and only biallelic SNPs were retained, an approach used previously in sugarcane (Zhang et al., 2018). In this method, the potential real states of autotetraploid heterozygous sites (such as CGGG, CCGG, or CCCG) were simplified to sites with two states (CG), and homozygous sites were not affected. This method therefore ignored the diversity of different real heterozygous sites (allelic dosage), and some analyses related to genetic diversity, such as the NJ analysis and nucleotide diversity, would have been affected. The NJ analysis is related to the diversity between individuals and may be greatly affected by ignoring allelic dosage, thereby producing inaccurate conclusions. To avoid misleading interpretations, the results of NJ analysis are not presented here. By contrast, GWAS is based primarily on the presence or absence of different genetic marker alleles in individuals, without regard to allelic dosage. Thus, the analyses presented here can overcome some challenges associated with autotetraploidy and can provide an example for future studies.

In this study, we report a highly continuous, chromosome-level reference genome of autotetraploid Zhongmu No. 1 alfalfa, as well as a population analysis of 162 worldwide alfalfa germplasms. These results provide insights into the genome evolution and the weak population structure of alfalfa. GWAS analysis identified a number of target regions that putatively control important agronomic traits. *MsFTa2*, which may be associated with several critical traits related to alfalfa's worldwide distribution, is an example of the value that target gene identification can bring to future functional studies and molecular breeding. These new genomic resources will promote the study of alfalfa genome evolution, perennation mechanisms, and other traits, thereby facilitating the breeding of improved forages to benefit human society.

METHODS

Plant Materials

The variety Zhongmu No. 1 (Chinese alfalfa) from the Institute of Animal Sciences of the Chinese Academy of Agricultural Sciences was chosen for the sequencing and assembly of the alfalfa reference genome. Zhongmu No. 1 was obtained by hybridizing several kinds of alfalfa. The accession is classified as *M. sativa* spp. *sativa*. Zhongmu No. 1 is a heterozygote with outcrossing that is largely self-incompatible, and it can therefore only be reproduced asexually (such as through cuttage). After germination, seeds were transferred to mixed soil (soil:vermiculite = 2:5) with MS medium (Murashige and Skoog, 1962). After growing for one month, one individual plant was selected for DNA extraction from leaves for SMRT PacBio, 10X genomics, and NGS sequencing. The selected plant was propagated asexually, and its progeny were grown to obtain more young leaf material for a Bionano optical map and Hi-C sequencing. The cutting seedling was transplanted onto mixed soil with MS medium.

Three weeks after cuttage, young emerging leaves were harvested. For RNA sampling, RNA was extracted from the following greenhouse-grown samples (Supplemental Table 7): stems without branches, stems with branches, leaf buds (three duplicates), flowers, flower buds, lateral roots, taproots, capsules, tillering area, young leaves (three duplicates), climax leaves (three duplicates), and young leaves after clipping. These samples were obtained from the same plant materials described in the PacBio protocol. RNA was also isolated from seeds, 21 dpi (days post infection) nodules, and 28 dpi nodules, which were planted on 13 × 13-cm dishes with MS medium. All of these fresh tissues were harvested, immediately frozen in liquid nitrogen, and then stored at −80°C until extraction.

For the soil treatments, plants were grown on dishes in well-watered conditions in a growth chamber at 24°C and 16 h daily light. After the plants had grown for 7 d, we inoculated them with *Sm1021* to obtain nodules. For resequencing samples, 137 species of *M. sativa* spp. *sativa* were collected from the United States Department of Agriculture Agricultural Research Service. These accessions were planted in soil for 2 weeks, and then young leaves were used for DNA extraction and sequencing.

DNA and RNA Isolation

Preparation of Genomic DNA for NGS Sequencing, 10X Genomics, and PacBio Sequencing

DNA was extracted from the young leaves of Zhongmu No. 1 with the modified cetyltrimethylammonium bromide (CTAB) method (Doyle, 1987). DNA samples were sent to BGI-Shenzhen (Shenzhen, China) and Wuhan Institute of Biotechnology (Wuhan, HuBei, China) for library construction and sequencing on the Illumina HiSeq 2000 platform (Illumina, San Diego, CA) and the PacBio Sequel platform (Pacific Biosciences, Menlo Park, CA).

Preparation for Hi-C

The Hi-C library was prepared and sequenced by ANOROAD-Beijing. In brief, chromatin in the nucleus of young alfalfa seedlings was fixed with formaldehyde and extracted. Fixed chromatin was digested with MboI, and sticky ends were filled in with biotinylated nucleotides and ligated. Purified DNA was treated to remove biotin.

Preparation for Bionano Optical Map

For Bionano mapping, young leaves of Zhongmu No. 1 were collected from the greenhouse, and high-molecular-weight (HMW) DNA was extracted following the Bionano IrysPrep High Polysaccharides Plant Tissue DNA Isolation User Guide (Bionano document no. 30128).

RNA Isolation for Transcriptome Sequencing

Samples were harvested into liquid nitrogen, then stored at −80°C until extraction. We used a high-throughput tissue grinder to grind the tissues. TRIzol (Invitrogen) and $CHCl_3$ were added to extract the RNA. RNA was washed with 75% ethanol and then dissolved in 30 μl RNase free water. The RNA samples were sent to BGI-Shenzhen for both RNA-seq and PacBio Iso-Seq.

Preparation for the Resequencing of Alfalfa Germplasms

Young leaves of each core germplasm were collected for genome sequencing. Genomic DNA was extracted using the CTAB method. DNA samples were sent to BGI-Shenzhen and Wuhan Institute of Biotechnology for library construction and sequencing on the BGISEQ-500 platform (BGI, Shenzhen, China).

Genome Sequencing

NGS Short Read Sequencing

The NGS data were sequenced by BGI using a BGISEQ-500. A total of 1 μg genomic DNA was randomly fragmented by ultrasound (Covaris). The fragmented genomic DNA was recycled with an average size of 200–400 bp using an Agencourt AMPure XP Medium kit. Fragments were end-repaired and then 3′ adenylated. Adaptors were ligated to the ends of the 3′ adenylated fragments, and fragments with adaptors were amplified by PCR. The double-stranded PCR products were heat-denatured and circularized by the splint oligo sequence. The single-stranded circular DNA molecules (ssCir DNA) were formatted as the final library. Quality-checked libraries were sequenced by BGISEQ-500. Each ssCir DNA molecule formed a DNA nanoball (DNB) that contained more than 300 copies through rolling-cycle replication. The DNBs were loaded into the patterned nanoarray using the high-density DNA nanochip technology. Finally, paired-end 100 or 150 bp reads were obtained by combinatorial probe-anchor synthesis. For a survey to estimate the genome, a total of 150.52 Gb clean data was generated. A total of 528.62 Gb clean data was generated to refine the assembled genome, and a total of 7.7 Tb clean data was generated for resequencing (Supplemental Data 2).

PacBio Sequencing

To enable an assembly of the complex alfalfa reference genome, we extracted DNA from the shoot tissue of alfalfa propagated by cuttage for PacBio sequencing. A total of 35 SMRT cells were run on the PacBio Sequel system by BGI. These cells generated 245.62 Gb of long read data with an N50 of 12 165 kb (Supplemental Table 2).

Bionano Optical Maps

Extracted HMW DNA molecules were fluorescently stained using Nick, Label, Repair, and Stain (NLRS) enzymatic reactions following the Bionano Prep Labeling - NLRS Protocol (Bionano document no. 30024). In brief, single-strand breaks were introduced into DNA molecules with the nicking enzyme Nb.BssSI (New England Biolabs) to generate sequence motif-specific patterns. Nicked sites were labeled with fluorescent nucleotides and repaired. Molecule backbones were also fluorescently stained with YOYO-1 to visualize their full lengths. NLRS reaction products were then run on the Bionano Saphyr system (BGI), in which the DNA molecules were automatically stretched and imaged within nanochannel arrays. Distances between fluorescently labeled nicking sites formed patterns that were used for alignment and assembly with Bionano Auto Detect software (version 2.1.4).

Hi-C Library Preparation and Sequencing

The Hi-C sequencing libraries were constructed by 10–12 cycles of PCR. Biotin-containing fragments were enriched using streptavidin C1 magnetic beads before the PCR amplification of the library. The library was sequenced on an Illumina HiSeq X platform (Illumina). The sequencing interacting pattern was obtained using an Illumina HiSeq instrument with 2 × 150-bp reads.

RNA-Seq

For RNA-seq, stranded RNA-seq libraries were constructed for each sample and sequenced on the BGISEQ-500 platform (BGI). For PacBio Iso-Seq, equal amounts of RNA from different tissues were pooled and then sequenced on the PacBio Sequel platform.

Genome Assembly

Genome Size Estimation

K-mer analysis was used to estimate the size and heterozygosity of the genome following the method described in Liu et al. (2013). We also used flow cytometry to estimate genome size. Following Dolezel et al. (2007), we used *G. max* and *Zea mays* as internal references (Supplemental Figure 3A–3D). As a control, the PI peak positions of Zhongmu No. 1, *G. max*, and *Z. mays* were estimated with flow cytometry (BD FACSCalibur). The samples were then mixed to make a nucleic solution to estimate the relative genome size of Zhongmu No. 1 (Supplemental Figure 3E and 3F).

Genome Assembly

The PacBio data were corrected with Canu (version 1.8 with useGrid = true; minThreads = 4; genomeSize = 1600m; minOverlapLength = 700; minReadLength = 1000; and other parameters set to default values). Next, we used MECAT to assemble the Canu-corrected reads with genomeSize = 1.6g; ErrorRate = 0.04; maxMemory = 500; maxThreads = 40; useGrid = 0; and Overlapper = mecat2asmpw. We used HERA (Du and Liang, 2019) to extend and connect the contigs and to fill in gaps in the Bionano hybrid scaffolds (Supplemental Figure 3). The self-alignment of the whole-genome contig sequences was performed using default the parameters of BWA-MEM, and the genome sequences were filtered with Redundans (with -t 10, –identity 0.55, –overlap 0.80, —noscaffolding, and –nogapclosing) and Purge Haplotigs (with default parameters). The overlap between sequences was merged using the results of BWA-MEM self-alignment, and the removed overlapping sequences were stored in a set of heterozygous sequences. Hi-C sequencing data were used to cluster the remaining genome sequences, resulting in the final 8 pseudo-chromosomes with a total length of 816 Mb and a contig N50 size of 3.9 Mb.

Refining the Genome

The NGS data were mapped to the genome using BWA-MEM (version 0.7.17) (Li and Durbin, 2009), and the results were filtered with Q30 by

SAMtools (version 1.8) (Li et al., 2009a, 2009b). Finally, the genome was corrected using Pilon (version 1.22) (Walker et al., 2014) based on the filtered alignments. Three rounds of genome correction were performed by Pilon.

Quality Control of the Assembly

We aligned the NGS data to the final pseudo-chromosomes with BWA-MEM. BUSCO (version 3, embryophyta_odb9 dataset) was used to evaluate the genome integrity.

Genome Annotation

Annotation of Repeat DNA Sequences

A combination of *ab initio* and homology-based methods was used to annotate repeats in the alfalfa genome. First, we constructed an *ab initio* repeat library using LTR_FINDER (version 1.05) (Xu and Wang, 2007) and RepeatModeler (version 1.0.1, see URLs) with default parameters. The predicted repeat library was aligned with the PGSB repeater database (see URLs) to separate repeats into distinct repeat families. Next, a repeat of the database from scratch and Repbase (version 20.11, see URLs) were input to RepeatMasker (version 4.0.7, see URLs) for alfalfa genomic execution based on a homologous repeat search. In addition, RepeatProteinMask was used with the wu-blastx search engine to identify any repeatability-related proteins that had been missed in the previous step. Finally, overlapping repeat sequences that belonged to the same repeat sequence were combined according to their compatibility in the genome. For overlapping duplicates that belonged to different repeat classes, the overlapping areas were divided in the middle. In addition, we used Tandem Repeats Finder (Benson, 1999) to annotate tandem repeats.

Prediction and Functional Annotation of Protein-Coding Genes

The RNA-seq data were aligned to the *M. sativa* genome using HISAT2 (version 2.1.0) (Kim et al., 2015) with default parameters, and transcripts were assembled using StringTie (version 1.3.5) (Pertea et al., 2016) with default parameters to obtain the transcript gff file. The genome was annotated to obtain reliable genes using the SwissProt (2019) protein database and MAKER gene annotation software. These genes were then used to train Augustus (version 3.2.3) (Keller et al., 2011) and SNAP (version 2006-07-28) (Korf, 2004) models. The transcripts of the RNA-seq data and the isoform sequences produced by PacBio sequencing were used as EST evidence, and the protein sequences of *Oryza sativa*, *Arabidopsis thaliana*, *Cicer arietinum*, *G. max*, *Medicago truncatula*, and SwissProt were used as protein evidence. Using the models trained by SNAP and Augustus, the second round of gene annotation was performed. The protein sequences of the obtained genes were annotated using InterProScan 5.0 (Jones et al., 2014). Based on the results of functional annotation and alignment with the TESeeker database (Kennedy et al., 2011), TE-related genes were filtered out.

Next, reliability classification of the genes was performed: the CDSs of the remaining genes were compared with the isoforms, and genes with identity >0.9 and coverage >0.75 were defined as high-confidence genes. The GC content, gene density, and LTR density were visualized using Circos (version 0.69-5) (Krzywinski et al., 2009) as shown in Figure 1. The GC content in a 500-kb window was scanned sequentially along the genome; the minimum value shown in the Circos plot is 0.3, and the maximum value is 0.4. The gene density in a 500-kb window was also scanned sequentially along the genome; the minimum value shown in the Circos plot is 0, and the maximum value is 35.

Genome Collinearity Analysis

The collinearity analysis of alfalfa and *M. truncatula* genes was performed using MCscan (see URLs) with the following command: python -m jcvi.compara.catalog ortholog –no_strip_names –nostdpf –cscore=.75. The result was filtered and integrated with the command: python -m jcvi.compara.synteny mcscan bed_file anchor_file –iter = 1.

Evolutionary Analysis

Based on the results of the collinearity analysis above, the synonymous substitutions per synonymous site (K_s) of the genes were calculated. The peak value was extracted using the K_s distribution map, and the differentiation time was estimated as Divergence time = $K_s/13 \times 1000$ (Mya) (Gaut et al., 1996).

Population Analysis

Genome Ploidy Estimation

Following Dolezel et al. (2007), we used Zhongmu No. 1 as a reference to estimate the PI peak position of the alfalfa samples, then estimated their ploidy for subsequent analysis.

Population Genome Resequencing and Detection of Nucleotide Variants

The whole-genome sequencing data of 137 alfalfa and 25 ssp. *caerulea* samples were mapped to the Zhongmu No. 1 genome with BWA (version 0.7.17) using the mem function. SAMtools (version 1.8) was used to filter the unmapped reads and sort the mapped reads. Additional filtration was performed with an in-house Perl script, and reads with mapping quality greater than 10 and the best alignment score higher than the second-best one were retained for further analysis. Nucleotide variants were then obtained using the Unified Genotyper tool of the Genome Analysis Toolkit (GATK) (version 3.4-46) (McKenna et al., 2010) with the following parameters: -stand_call_conf 50.0, -stand_emit_conf 10.0, -dcov 1000, -A Coverage, and -A AlleleBalance. The raw variants were filtered using the GATK VariantFiltration tool. The criteria used to filter the raw variants were QUAL <50.0, MQ0 $\geq$ 4 && ((MQ0/(1.0*DP)) > 0.1, DP < 5, QD < 1.5, and clusterWindowSize = 10. After filtration, we obtained 116 851 522 SNPs and InDels, including 11 827 328 biallelic SNPs. All nucleotide variants were annotated for their potential impacts on coding genes using snpEff (version 3.3) (Cingolani et al., 2012).

Population Structure and Phylogenetic Analysis

From the obtained nucleotide variants, 986 141 SNP sites with two allelic types that fell within annotated CDS regions and were missing in less than 5% of the population were used for the population structure analysis of 163 *M. sativa* samples. Population structure was inferred using ADMIXTURE (version 1.23) (Alexander et al., 2009) with K from 2 to 5, each with 200 bootstraps. PCAs were performed using the smartpca function in EIGENSOFT (version 5.0.2) (Patterson et al., 2006). Pairwise differences for all samples were also calculated on these SNPs with an in-house Perl script. The genotypes of the SNPs were transformed into multi-consensus fasta format, and neighbor-joining trees were constructed with the R package “ape.” Heterozygous genotypes were transformed into four forms for neighbor-joining tree construction: missing, major allele, minor allele, and random allele. The same SNP dataset was transformed into a population allele frequency table based on the sample grouping result of ADMIXTURE when K = 3, and population split and mixture analyses were performed with TreeMix (version 1.13) (Pickrell and Pritchard, 2012). A subset of 322 859 SNPs with a minor allele frequency over 0.05 in 127 relatively clear tetraploid alfalfa samples were used for population structure inference and PCA as described above.

Nucleotide Diversity and Fixation Analysis

For each population separated according to the population structure analysis, nucleotide diversities (π) were calculated using VCFtools (version 0.1.15) (Danecek et al., 2011) with a 100-kb sliding window and a 100-kb step length. Fixation indices between each population were also calculated using VCFtools with a 100-kb sliding window and a 100-kb step length.

GWAS Analysis

From the total set of nucleotide variants, 2 896 472 biallelic SNPs with a missing rate of less than 0.25 and a minor allele frequency greater than 0.05 were selected for GWAS analysis using EMMAX (version 20120210) (Kang et al., 2010). To determine the genome-wide significance thresholds, permutation tests were performed 200 times for each trait, and the 10% lowest tail from the 200 recorded minimal *P* values (false discovery rate < 10%) was taken as the threshold. The results of EMMAX were visualized as Manhattan and Q-Q plots with the R package "qqman" (Turner, 2014) and in-house R scripts based on the package "ggplot2" (Wickham, 2016). We then selected genes near the peak SNPs according to the Manhattan plot (Supplementary Figure 10) for each trait and BLASTed them against *A. thaliana* to obtain gene names listed in Data S3. Spearman's rank correlation coefficient (rho) was calculated among the SNPs.

Salt Treatment and Gene Expression Analysis

To examine the expression patterns of the *FT*s, *M. sativa* accessions were cultured on MS medium with or without 100 mM NaCl in petri dishes under 16 h light (200 μmol m^{-2} s^{-1})/8 h dark conditions and 70% relative humidity for one week. First, the copy numbers of the *FT*s and housekeeping genes were determined by qPCR as described previously (D'Haene et al., 2010). The relative expression of the *FT*s was measured by qRT–PCR using a CFX-96 Real-Time System (Bio-Rad) and the SYBR Premix Ex Taq (TaKaRa, RR420A). RNA was extracted with the TRIzol reagent (Ambion, 15596018), and cDNA was obtained by reverse transcription using the M-MLV reverse transcriptase (Promega, M1701). The relative expression of the target genes was normalized to that of *Actin*, *EF1a*, and *EIF4A*. The primers used were FTa1-cnv-F: TTCCTCTCCGAGTGACCTATG; FTa1-cnv-R: ATCGTTTCCACCAACACTCA; FTa2-cnv-F: CTTGCTGTTGGGCGTGTA; FTa2-cnv-R: GGGAAGGTTTAAGCTCACGA; FTc-cnv-F: TTCATCGATTTGTGATTGCAT; FTc-cnv-R: GCTCTCTTTGGCAGTTGAAA; Actin-cnv-F: CCATTGAGCACGGTATTGTC; Actin-cnv-R: ATTGGCCTTTGGGTTAAGTG; EF1a-cnv-F: TGCCTTGTGGAAATTTGAGA; EF1a-cnv-R: AGCCTGGGAGGTTCCAGTA; EIF4A-cnv-F: GCTCTGGCTCTTCTCGAGTT; EIF4A-cnv-R: TTCAGGTTGGGTAGGCAAAT; FTa1-qpcr-F: ATGGCCGGTAGCAGTAGGAATC; FTa1-qpcr-R: AAAGTGGGGTTACTTGGGCT; FTa2-qpcr-F: TGACTGATATTCCAGCAACTAATG; FTa2-qpcr-R: CGGTGATCCAAGATCGTAAAA; FTc-qpcr-F: CGGATATTCCAGCAACAACAA; FTc-qpcr-R: CATCTCCTTCCACCGCAAC; Actin4A-RT-F: CCAAAGGCCAACAGAGAAAA; Actin4A-RT-R: ACGACCAGCAAGATCCAAAC; EF1a-RT-F: GCACGCTCTTCTTGCCTTTA; EF1a-RT-R: GTCACCTTCAAATCCGGAGA; EIF4A-RT-F: TTTAGCTCCGGAAGGTTCAC; EIF4A-RT-R: TGCTGAATCACATCGAGACC.

URLs

RepeatModeler, http://repeatmasker.org/RepeatModeler/; PGSB repeater database, http://pgsb.helmholtz-muenchen.de/plant/recat/; Repbase http://www.girinst.org/repbase; RepeatMasker, http://www.repeatmasker.org; MCscan, https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version).

ACCESSION NUMBERS

All raw data were uploaded to the National Genomics Data Center (https://bigd.big.ac.cn/) under BioProject PRJCA001722. Genome assembly and gene annotation files are available at https://figshare.com/articles/dataset/Medicago_sativa_genome_and_annotation_files/12623960.

SUPPLEMENTAL INFORMATION

Supplemental Information is available at *Molecular Plant Online*.

FUNDING

This work was supported by the National Key Research & Development Program of China (2019YFD1002701), Agricultural Variety Improvemnt Project of Shandong Province (2019LZGC010), and the Project for Extramural Scientists of the State Key Laboratory for Agrobiotechnology (2020SKLAB6-15, 2011SKLAB01-1).

AUTHOR CONTRIBUTIONS

T.W. and J.D. designed the research. T.W., C.L., and J.D. supervised the research. C.S., H.D., Z.C., H.L., J.D., C.L., and T.W. wrote the paper. H.D. and C.S. performed the genome assembly. H.L., C.S., X.L., and H.D. performed the genome annotation and evolution analysis. Z.C., C.S., F.Z., and P.L. performed the population structure and GWAS analysis. C.S., H.C., X.M., Q.L., and L.Z. performed the experiments.

ACKNOWLEDGMENTS

We thank Qingchuan Yang for providing the seeds of Zhongmu No. 1 and NPGS-ARS for providing resequence accessions. We thank Jinsheng Lai and Junpeng Shi for their help in revising the manuscript. No conflict of interest declared.

Received: April 20, 2020
Revised: June 24, 2020
Accepted: July 10, 2020
Published: July 13, 2020

REFERENCES

Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. **19**:1655–1664.

Annicchiarico, P., Nazzicari, N., and Brummer, E. (2016). Alfalfa genomic selection: challenges, strategies, transnational cooperation. In Breeding in a World of Scarcity, C.H. Cham, ed. (Switzerland: Springer), pp. 145–149.

Barros, J., Temple, S., and Dixon, R.A. (2019). Development and commercialization of reduced lignin alfalfa. Curr. Opin. Biotechnol. **56**:48–54.

Basigalup, D., Barnes, D., and Stucker, R. (1995). Development of a core collection for perennial *Medicago* plant introductions. Crop Sci. **35**:1163–1168.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. **27**:573–580.

Bhandari, H.S., Pierce, C.A., Murray, L.W., and Ray, I.M. (2007). Combining abilities and heterosis for forage yield among high-yielding accessions of the alfalfa core collection. Crop Sci. **47**. https://doi.org/10.2135/cropsci2006.06.0398.

Butler, A.J.A.P. (1995). The small-seeded legumes: an enigmatic prehistoric resource. Acta Palaeobotanica **35**:105–115.

Cao, D., Osborn, T.C., and Doerge, R.W. (2004). Correct estimation of preferential chromosome pairing in autotetraploids. Genome Res. **14**:459–462.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly **6**:80–92.

Cooke, J.E., Eriksson, M.E., and Junttila, O. (2012). The dynamic nature of bud dormancy in trees: environmental control and molecular mechanisms. Plant Cell Environ. **35**:1707–1728.

D'Haene, B., Vandesompele, J., and Hellemans, J. (2010). Accurate and objective copy number profiling using real-time quantitative PCR. Methods **50**:262–270.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics **27**:2156–2158.

Dekker, J. (2006). The three 'C's of chromosome conformation capture: controls, controls, controls. Nat. Methods **3**:17.

Dolezel, J., Greilhuber, J., and Suda, J. (2007). Estimation of nuclear DNA content in plants using flow cytometry. Nat. Protoc. **2**:2233–2244.

Doyle, J.J. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. **19**:11–15.

Du, H., and Liang, C. (2019). Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads. Nat. Commun. **10**:5360.

Edger, P.P., Poorten, T.J., VanBuren, R., Hardigan, M.A., Colle, M., McKain, M.R., Smith, R.D., Teresi, S.J., Nelson, A.D.L., Wai, C.M., et al. (2019). Origin and evolution of the octoploid strawberry genome. Nat. Genet. **51**:541–547.

Elfaki, M.O., and Abdelatti, K.A. (2016). Rumen content as animal feed: a review. U. K. J. Vet. Med. Anim. Prod. **7**:80–88.

F Blondon, D.M., Brown, S., and Kondorosi, A. (1994). Genome size and base composition in *Medicago sativa* and *M. truncatula* species. Genome **37**:264–270.

Fyad-Lameche, F.-Z., Iantcheva, A., Siljak-Yakovlev, S., and Brown, S.C. (2015). Chromosome number, genome size, seed storage protein profile and competence for direct somatic embryo formation in Algerian annual *Medicago* species. Plant Cell Tissue Organ Cult. **124**:531–540.

Gaut, B.S., Morton, B.R., McCaig, B.C., and Clegg, M.T. (1996). Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. Proc. Natl. Acad. Sci. U S A **93**:10274–10279.

Hanson, A.A., Barnes, D., Hill, R.R., Heichel, G.H., Leath, K., Hunt, O., Marten, G., Tesar, M., Mickelson, S., and Holtgraver, K. (1988). Alfalfa and Alfalfa Improvement (Madison, WI: American Society of Agronomy).

Hawkins, C., and Yu, L.-X. (2018). Recent progress in alfalfa (*Medicago sativa* L.) genomics and genomic selection. Crop J. **6**:565–575.

Havananda, T. (2010). Relationships among diploid members of the *Medicago sativa* (Fabaceae) species complex based on chloroplast and mitochondrial DNA sequences[J]. Syst. Bot. **35** (1):140–150.

Jia, C., Wu, X., Chen, M., Wang, Y., Liu, X., Gong, P., Xu, Q., Wang, X., Gao, H., and Wang, Z. (2017). Identification of genetic loci associated with crude protein and mineral concentrations in alfalfa (*Medicago sativa*) using association mapping. BMC Plant Biol. **17**:97.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics **30**:1236–1240.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. **42**:348–354.

Kardailsky, I., Shukla, V.K., Ahn, J.H., Dagenais, N., Christensen, S.K., Nguyen, J.T., Chory, J., Harrison, M.J., and Weigel, D. (1999). Activation tagging of the floral inducer FT. Science **286**:1962–1965.

Keller, O., Kollmar, M., Stanke, M., and Waack, S. (2011). A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics **27**:757–763.

Kennedy, R.C., Unger, M.F., Christley, S., Collins, F.H., and Madey, G.R. (2011). An automated homology-based approach for identifying transposable elements. BMC Bioinformatics **12**:130.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods **12**:357–360.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. **27**:722–736.

Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics **5**:59.

Kreplak, J., Madoui, M.A., Cápal, P., Novák, P., Labadie, K., Aubert, G., Bayer, P.E., Gali, K.K., Syme, R.A., Main, D., et al. (2019). A reference genome for pea provides insight into legume genome evolution. Nat. Genet. **51**:1411–1422.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res. **19**:1639–1645.

Lam, E.T., Hastie, A., Lin, C., Ehrlich, D., Das, S.K., Austin, M.D., Deshpande, P., Cao, H., Nagarajan, N., Xiao, M., et al. (2012). Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat. Biotechnol. **30**:771–776.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**:1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009a). The sequence alignment/map format and SAMtools. Bioinformatics **25**:2078–2079.

Li, X., Wei, Y., Nettleton, D., and Brummer, E.C. (2009b). Comparative gene expression profiles between heterotic and non-heterotic hybrids of tetraploid *Medicago sativa*. BMC Plant Biol. **9**:107.

Li, X., Wei, Y., Acharya, A., Jiang, Q., Kang, J., and Brummer, E.C. (2014). A saturated genetic linkage map of autotetraploid alfalfa (*Medicago sativa* L.) developed using genotyping-by-sequencing is highly syntenous with the *Medicago truncatula* genome. G3 (Bethesda) **4**:1971–1979.

Li, C., Cai, D.R., Gu, C., and Huang, X.Z. (2015). Gossypium barbadense GbMFT2 can improve salt-resistant capability in tobacco, Fenzi Zhiwu Yuzhong. Mol. Plant Breed. **13**:1009–1016.

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., Li, Z., Chen, Y., Mu, D., and Fan, W. (2013). Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome project. arXiv, 1308.2012. https://arxiv.org/abs/1308.2012.

Liu, Z., Li, X., Wang, Z., and Sun, Q. (2015). Contrasting strategies of alfalfa stem elongation in response to fall dormancy in early growth stage: the tradeoff between internode length and internode number. PLoS One **10**:e0135934.

Liu, X.P., Hawkins, C., Peel, M.D., and Yu, L.X. (2019a). Genetic loci associated with salt tolerance in advanced breeding populations of tetraploid alfalfa using genome-wide association studies. Plant Genome **12**. https://doi.org/10.3835/plantgenome2018.05.0026.

Liu, Z.Y., Baoyin, T., Li, X.L., and Wang, Z.L. (2019b). How fall dormancy benefits alfalfa winter-survival? Physiologic and transcriptomic analyses of dormancy process. BMC Plant Biol. **19**:205.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. **20**:1297–1303.

Michaud, R., Lehman, W., and Rumbaugh, M. (1988). World distribution and historical development. In Alfalfa and Alfalfa Improvement (Madison, WI: American Society of Agronomy), pp. 25–91.

Murashige, T., and Skoog, F.J.P.p. (1962). A revised medium for rapid growth and bio assays with tobacco tissue cultures. Physiol. Plant. **15**:473–497.

Ou, S., Chen, J., and Jiang, N. (2018). Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res. gky730. https://doi.org/10.1093/nar/gky730.

Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. Plos Genet. **2**:e190.

Pecrix, Y., Staton, S.E., Sallet, E., Lelandais-Briere, C., Moreau, S., Carrere, S., Blein, T., Jardinaud, M.F., Latrasse, D., Zouine, M., et al. (2018). Whole-genome landscape of *Medicago truncatula* symbiotic genes. Nat. Plants **4**:1017–1025.

Pertea, M., Kim, D., Pertea, G.M., Leek, J.T., and Salzberg, S.L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat. Protoc. **11**:1650–1667.

Pickrell, J.K., and Pritchard, J.K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. Plos Genet. **8**:e1002967.

Pryszcz, L.P., and Gabaldón, T. (2016). Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res. **44**:e113.

Radović, J., Sokolović, D., and Marković, J. (2009). Alfalfa—most important perennial forage legume in animal husbandry. Biotechnol. Anim. Husbandry **25**:465–475.

Roach, M.J., Schmidt, S.A., and Borneman, A.R. (2018). Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics **19**:460.

Russelle, M.P. (2014). Alfalfa: after an 8,000-year journey, the "Queen of Forages" stands poised to enjoy renewed popularity. Am. Scientist **89**:252–261.

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Kato, T., Nakao, M., Sasamoto, S., Watanabe, A., Ono, A., Kawashima, K., et al. (2008). Genome structure of the legume, *Lotus japonicus*. DNA Res. **15**:227–239.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., et al. (2010). Genome sequence of the palaeopolyploid soybean. Nature **463**:178–183.

Stanton-Geddes, J., Paape, T., Epstein, B., Briskine, R., Yoder, J., Mudge, J., Bharti, A.K., Farmer, A.D., Zhou, P., Denny, R., et al. (2013). Candidate genes and genetic architecture of symbiotic and agronomic traits revealed by whole-genome, sequence-based association genetics in *Medicago truncatula*. PLoS One **8**:e65688.

Tamaki, S., Matsuo, S., Wong, H.L., Yokoi, S., and Shimamoto, K. (2007). Hd3a protein is a mobile flowering signal in rice. Science **316**:1033–1036.

Turner, S.D. (2014). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots[J]. bioRxiv https://doi.org/10.1101/005165.

Tysdal, H.M., Kiesselbach, T.A., and Westover, H.M. (1942). Alfalfa breeding (Research Bulletin: Bulletin of the Agricultural Experiment Station of Nebraska No. 124). https://digitalcommons.unl.edu/ardhistrb/219/.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., and Young, S.K. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One **9**:e112963.

Waterhouse, R.M., Seppey, M., Simao, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., and Zdobnov, E.M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. https://doi.org/10.1093/molbev/msx319.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (New York, NY: Springer-Verlag New York).

Xiao, C.L., Chen, Y., Xie, S.Q., Chen, K.N., Wang, Y., Han, Y., Luo, F., and Xie, Z. (2017). MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. Nat. Methods **14**:1072–1074.

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. **35**:W265–W268.

Yu, F., Wang, H., Zhao, Y., Liu, R., Dou, Q., Dong, J., and Wang, T. (2017). Karyotypic evolution of the *Medicago* complex: sativa-caerulea-falcata inferred from comparative cytogenetic analysis. BMC Evol. Biol. **17**:104.

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

Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., Zhu, F., Jones, T., Zhu, X., Bowers, J., et al. (2018). Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. Nat. Genet. **50**:1565–1573.