Plant Communications
Correspondence

CellPress
Partner Journal

The haplotype-resolved and near-telomere-to-telomere genome assembly for the autotetraploid alfalfa

Dear Editor,

Alfalfa (*Medicago sativa*) is a valuable forage crop, cultivated worldwide as feed for ruminants due to its rich protein content and high palatability. The release of chromosome-scale (Zhongmu-1, Xinjiang Daye) and allele-aware (Zhongmu-4) genome assemblies (Long et al., 2022) has greatly accelerated the discovery of molecular mechanisms and regulatory networks underlying agronomic traits, thereby promoting genetic improvements. However, substantial genomic gaps in the telomeric and centromeric regions remain, limiting the understanding of the full genomic context, functional gene discovery, and breeding progress.

Here, we report a near-complete, allele-aware genome assembly of alfalfa Zhongmu-4 (ZM4), achieved by integrating Oxford Nanopore Technologies (ONT) ultra-long reads, PacBio high-fidelity (HiFi) reads, and Pore-C sequencing data (Supplemental Table 1). We first generated a preliminary assembly using Hifiasm (Cheng et al., 2024), combining 34× HiFi reads (>15 kb) and 35× ONT ultra-long reads (>100 kb), which yielded a contig N50 of 15.12 Mb (Supplemental Table 2). Four haplotypes were subsequently phased using HapHiC (Zeng et al., 2024) guided by Pore-C signals. Structural accuracy was validated by the presence of clear diagonal signals in the Pore-C contact maps. To resolve the remaining gaps, a strategy leveraging the assembly graph structure (Bi et al., 2024) was employed to manually fill gaps along with the integration of additional contigs from alternative assemblies (Supplemental Table 3). To reconstruct telomeres, high-quality ONT reads were aligned to chromosomes lacking detectable telomeric sequences, enabling manual extension of 13 telomeres to their full lengths (Supplemental Table 4). Ultimately, these combined efforts resulted in a final genome assembly of 3136 Mb (Figures 1A and 1B), with only eight unresolved gaps remaining (Supplemental Table 5), representing the most complete alfalfa genome assembly to date.

To evaluate the quality of the genome assembly, we performed Benchmarking Universal Single-Copy Orthologs analysis, which detected 99.5% of expected orthologs (Supplemental Table 6), an improvement over existing alfalfa genome assemblies (93.7%–99.1%). Across the ZM4 assembly, Illumina, HiFi, and ONT reads exhibited stable mapping coverage, with alignment rates of 99.7%, 99.8%, and 97.9%, respectively (Figure 1C). Notably, we identified a large segment inversion (∼30 Mb) on chromosome 7_4. The inversion's structural accuracy was validated by ONT reads in the Integrative Genomics Viewer (Figure 1D; Supplemental Figure 1). Furthermore, genes in the vicinity of the inversion breakpoints exhibited altered expression (Supplemental Figure 2). The assembly achieved an overall quality value score of 58.4 and demonstrated retrotransposon completeness, with a long terminal repeat assembly index of 20.24 (Supplemental Figure 3). Collectively, these benchmark values meet the stringent criteria of a high-quality, near-complete genome assembly.

To generate high-quality gene annotation, we integrated evidence from *de novo* prediction, homology-based alignment, and transcriptome data from multiple tissues (leaf, root, flower, and stem) and stress treatments (drought, salt, and cold). This comprehensive approach identified a total of 193 260 genes (Supplemental Table 7). Benchmarking Universal Single-Copy Orthologs analysis further confirmed the completeness of the gene set, with 99.5% of conserved orthologs successfully recovered. Using the longest haplotype as a reference genome, we systematically characterized 48 498 protein-coding genes with clearly defined allelic variants across tetraploid chromosomal regions. A majority of genes (63.5%, 30 802 genes) exhibited tetra-allelic variation, followed by tri-allelic (13.0%, 6301 genes), bi-allelic (11.4%, 5528 genes), and mono-allelic (12.8%, 6220 genes) variation. Transposable element annotation revealed that repetitive sequences accounted for 70.39% of the genome. Long terminal repeat retrotransposons are predominant (42.46%), with *Gypsy/DIRS1* elements comprising 8.05% and *Ty1/Copia* elements comprising 4.42% of the genome.

Compared with the previous genome version, the current assembly is 580.16 Mb longer, with chromosome 7 showing the most substantial increase in sequence. In total, 50 726 new genes were identified (Figure 1E), with Gene Ontology terms associated with critical biological functions, including adaptation to light stress, symbiotic microbial interactions, and root nodulation (Supplemental Figure 4).

In this assembly, the telomeres and putative centromeres were accurately identified using Quartet, based on the repeat monomer sequence "AAACCCT" (Figure 1A). For the telomeres, repeat copy number varied substantially, ranging from 131 to 2175 copies among different chromosome ends. Centromeric regions, ranging from 5 197 602 to 10 654 325 bp in length, were characterized by relatively low GC content (Figure 1A; Supplemental Table 8). They were also enriched in monomeric repeats MsCR-1, MsCR-2, and MsCR-3 (Yu et al., 2013), which varied in proportion across different chromosomes (Figure 1A; Supplemental Figure 5). Furthermore, chr1_1, chr2_1–4, and chr3_3–4 putative centromeres were enriched in *Gypsy* elements (fraction $\geq$ 0.5). Additionally, substantial structural variations between homologous chromosomes were identified,

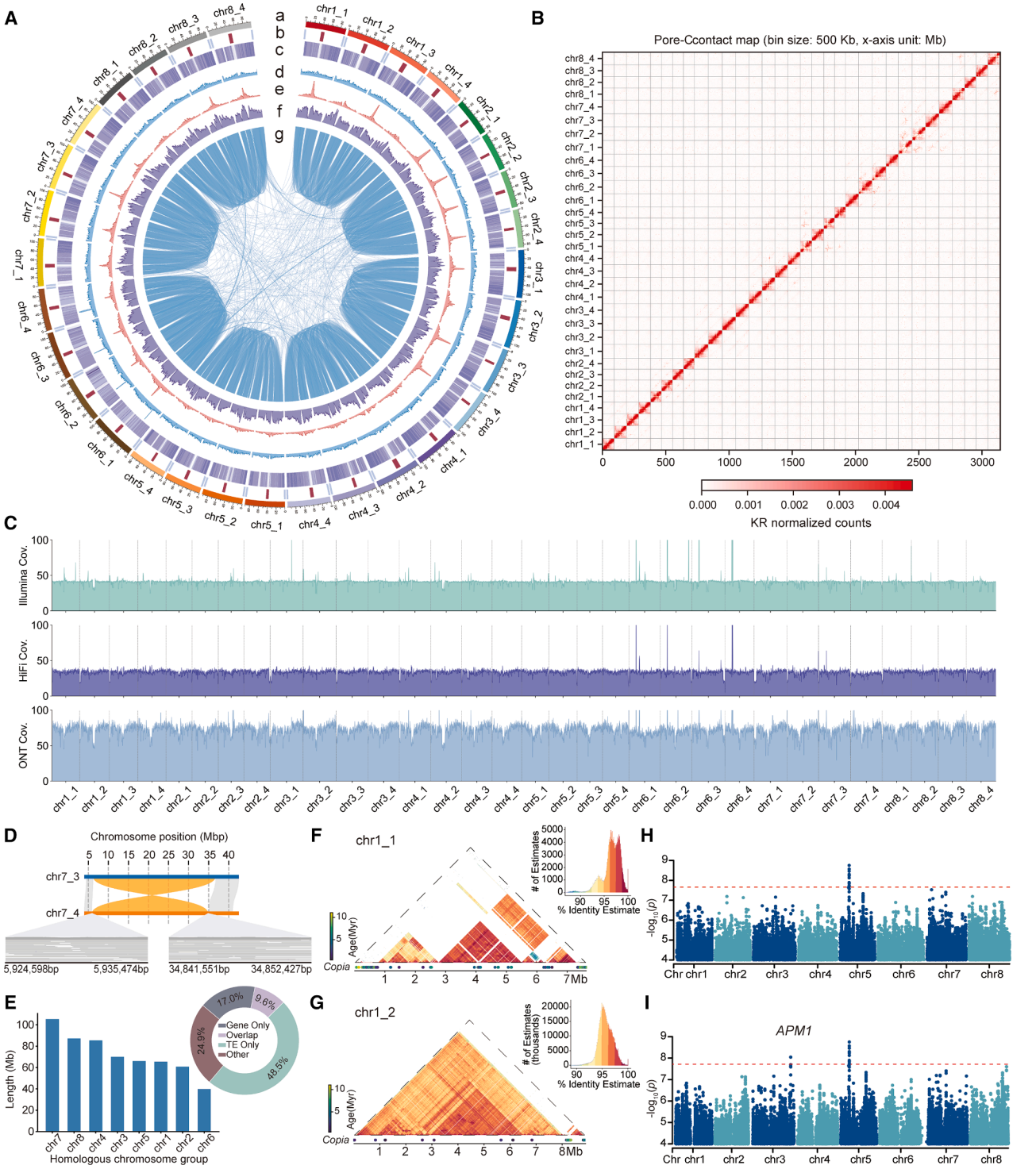

Figure 1. Comprehensive genomic characterization and validation of the ZM4 assembly.

(A) Circos plot displaying the genomic features of the ZM4 assembly. Tracks from outside to inside represent **(a)** chromosome length, **(b)** centromeric and telomeric positions, **(c)** gene density, **(d)** GC content density, **(e)** *Gypsy* long terminal repeat (LTR) retrotransposon density, **(f)** *Copia* LTR retrotransposon density, and **(g)** allelic genomic synteny relationships.

(B) Genome-wide Pore-C contact heatmap demonstrating chromosomal organization and structural accuracy.

(C) Genome-wide mapping coverage (Cov.) of Illumina, HiFi, and ONT reads.

(D) Validation of a large-segment inversion (~30 Mb) between chromosome 7 homologs (chr7_3 and chr7_4), confirmed by visualization of ONT read alignments in the Integrative Genomics Viewer.

(legend continued on next page)

primarily driven by *Copia* retrotransposon accumulation (Figures 1F and 1G). Specifically, clusters of older *Copia* elements within centromeric regions have undergone extensive mutations, deletions, and rearrangements over evolutionary time, resulting in reduced sequence similarity, decreased identity scores, and shorter alignable regions during self-alignment analyses, thereby increasing the structural complexity of these regions.

Based on variant calling of 176 clonally propagated, diverse alfalfa accessions against the near-complete genome assembly, and phenotypic data collected across multiple years (2022–2023) from two locations in Ningxia, China, we conducted genome-wide association studies (GWAS) to dissect alfalfa forage quality traits. We focused particularly on crude protein, a crucial trait directly related to animal performance. Identification of genes associated with leaf development is also of particular interest, given that leaves contain the majority of alfalfa protein (Hojilla-Evangelista et al., 2017). Comparing GWAS results between assembly versions (Figures 1H and 1I) revealed a new candidate gene, *Aminopeptidase M1* (*APM1*), which regulates auxin polar transport and metabolism and potentially controls leaf size (Peer et al., 2009; Teale and Palme, 2018). Consistent with this theory, *APM1* expression is significantly lower in the high crude protein group (Welch's *t*-test, $p < 0.01$), indicating a negative correlation between crude protein content and *APM1* expression (Supplemental Figure 6). Notably, the updated genome assembly enabled the identification of this gene, offering a promising lead for further exploration of its role in leaf development.

In conclusion, the near-complete genome assembly of ZM4 resolved gaps totaling 580.16 Mb, enabling the discovery of novel GWAS signals and candidate genes. This improved assembly unlocks the analysis of complex genomic regions, including gene clusters, repetitive elements, and putative centromeres and telomeres, thereby facilitating in-depth analyses of structural variation, genome evolution, and regulatory networks underlying key agronomic traits. This genomic resource establishes a robust foundation for precision breeding in alfalfa.

DATA AND CODE AVAILABILITY

Raw sequencing data are available at the National Genomics Data Center under BioProject accession PRJCA041059. The genome assembly and gene annotation files have been deposited in Figshare (https://doi.org/10.6084/m9.figshare.30768716).

FUNDING

This work was supported by the National Key R&D Program of China (2025YFF1000700), the Central Public-interest Scientific Institution Basal Research Fund (Y2025YC44), the Biological Breeding–National Science and Technology Major Project (2022ZD04011), the China Agriculture Research System of MOF and MARA (CARS-34), and the National Natural Science Foundation of China (32222019 and 32370714).

ACKNOWLEDGMENTS

No conflict of interest declared.

AUTHOR CONTRIBUTIONS

R.L., X. Zhang, Q.Y., and J.K. designed the study; B.S., X.J., F.H., X. Zeng, Y.W., and Y.Z. performed the statistical and bioinformatic analyses; M.L., X.W., T.Z., L.C., M.X., Y.X., K.Z., H.Z., and H.Y. conducted the field and laboratory experiments; B.S. and X.J. wrote the manuscript with guidance from R.L. and X. Zhang; F.H. revised the manuscript; and all authors read and approved the final manuscript.

SUPPLEMENTAL INFORMATION

Supplemental information is available at *Plant Communications Online*.

Received: June 30, 2025
Revised: October 24, 2025
Accepted: December 29, 2025
Published: December 30, 2025

Bilig Sod[1,2,3], Xueqian Jiang[1,3], Fei He[1,3], Xiaofei Zeng[2], Yibin Wang[2], Yaodong Zheng[2], Mingna Li[1], Xue Wang[1], Tiejun Zhang[1], Lin Chen[1], Ming Xu[1], Yanchao Xu[1], Kai Zhu[1], He Zhu[1], Haojie Yu[1], Junmei Kang[1,], Qingchuan Yang[1,*], Xingtan Zhang[2,*] and Ruicai Long[1,*]*

[1]Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing, China
[2]State Key Laboratory of Tropical Crop Breeding, Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China
[3]These authors contributed equally to this article.
***Correspondence: Junmei Kang (kangjunmei@caas.cn), Qingchuan Yang (yangqingchuan@caas.cn), Xingtan Zhang (zhangxingtan@caas.cn), Ruicai Long (longruicai@caas.cn)**
https://doi.org/10.1016/j.xplc.2025.101691

REFERENCES

Bi, G., Zhao, S., Yao, J., Wang, H., Zhao, M., Sun, Y., Hou, X., Haas, F. B., Varshney, D., Prigge, M., et al. (2024). Near telomere-to-telomere genome of the model plant Physcomitrium patens. Nat. Plants **10**:327–343.

Cheng, H., Asri, M., Lucas, J., Koren, S., and Li, H. (2024). Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. Nat. Methods **21**:967–970.

Hojilla-Evangelista, M.P., Selling, G.W., Hatfield, R., and Digman, M. (2017). Extraction, composition, and functional properties of dried alfalfa (Medicago sativa L.) leaf protein. J. Sci. Food Agric. **97**:882–888.

Long, R., Zhang, F., Zhang, Z., Li, M., Chen, L., Wang, X., Liu, W., Zhang, T., Yu, L.-X., He, F., et al. (2022). Genome Assembly of Alfalfa Cultivar Zhongmu-4 and Identification of SNPs Associated with Agronomic Traits. Genom. Proteom. Bioinform. **20**:14–28.

(E) Distribution of newly assembled genomic sequences (580.16 Mb) across chromosomes, highlighting chromosome 7 with the largest increase. The inset pie chart shows the annotation breakdown: transposable element (TE)-only regions (48.5%), gene-only regions (17.0%), regions containing both genes and TEs (9.6%), and unannotated sequences (other; 24.9%).
(F and G) Dot plots illustrating self-alignment of centromeric regions from chromosome 1 homologs (chr1_1 and chr1_2), revealing a correlation between the accumulation of older *Copia* clusters and reduced sequence identity and alignment completeness. Inset histograms show the distribution of sequence identity.
(H and I) Manhattan plots of crude protein content based on the previous **(H)** and current **(I)** genome assemblies.

Peer, W.A., Hosein, F.N., Bandyopadhyay, A., Makam, S.N., Otegui, M. S., Lee, G.-J., Blakeslee, J.J., Cheng, Y., Titapiwatanakun, B., Yakubov, B., et al. (2009). Mutation of the Membrane-Associated M1 Protease APM1 Results in Distinct Embryonic and Seedling Developmental Defects in Arabidopsis. Plant Cell **21**:1693–1721.

Teale, W., and Palme, K. (2018). Naphthylphthalamic acid and the mechanism of polar auxin transport. J. Exp. Bot. **69**:303–312.

Yu, F., Lei, Y., Li, Y., Dou, Q., Wang, H., and Chen, Z. (2013). Cloning and characterization of chromosomal markers in alfalfa (Medicago sativa L.). Theor. Appl. Genet. **126**:1885–1896.

Zeng, X., Yi, Z., Zhang, X., Du, Y., Li, Y., Zhou, Z., Chen, S., Zhao, H., Yang, S., Wang, Y., and Chen, G. (2024). Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. Nat. Plants **10**:1184–1200.