bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

A subgenome-resolved and chromosome-scale reference genome assembly of allotetraploid wheat wild relative *Aegilops peregrina*

Jatinder Singh[1,3], Santosh Gudi[1,3], Peter J. Maughan[2], Upinder Gill[1], and Rajeev Gupta[3*]

[1]Department of Plant Pathology, Microbiology, and Biotechnology, North Dakota State University, Fargo, ND, USA

[2]Brigham Young University, Provo, UT, USA

[3]Cereal Crops Improvement Research Unit, Edward T. Schafer Agricultural Research Center, USDA-ARS, Fargo, ND, USA

*Corresponding author (email: rajeev.gupta@usda.gov)

ORCID

Jatinder Singh: 0000-0002-0873-3677

Santosh Gudi: 0000-0003-1885-8139

Peter J. Maughan: 0000-0003-3714-3411

Upinder Gill: 0000-0003-4572-5243

Rajeev Gupta: 0000-0002-1451-215X

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Abstract

Aegilops peregrina is a wild allotetraploid wheat wild relative and an important source of genetic diversity for stress tolerance and agronomic traits. Here, we report a subgenome-resolved, chromosome-scale reference genome assembly of a drought tolerant and stem rust resistant *Ae. peregrina* accession PI 604178 generated using PacBio HiFi and Hi-C sequencing. The 10.13 Gb assembly contains 98.81% of sequence anchored to 14 pseudomolecules representing the seven S^p and seven U^p chromosomes, with contig and scaffold N50 values of 25.84 and 746.48 Mb, respectively. The assembly achieved a consensus quality value of 74.61, 97.83% k-mers completeness, and 99.9% BUSCO completeness. LTR Assembly Index values of 20.43 and 18.79 for the S^p and U^p subgenomes, respectively, further supported high continuity across repeat-rich regions. Repetitive elements comprise 85.93% of chromosome-anchored assembly. We annotated 59,910 high-confidence protein-coding genes, with comparable gene representation across the two subgenomes. This reference genome provides a high-quality genomic framework for comparative analyses, characterization of important loci regulating agronomic and resilience related traits, and sequence-guided exploitation of *Ae. peregrina* allelic diversity for wheat improvement.

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Background & Summary

Bread wheat (*Triticum aestivum* L.) provides approximately one-fifth of the calories consumed globally[1]. But its genetic diversity is constrained by the limited number of hybridization events involved in the origin of hexaploid wheat[2–4]. Wheat wild relatives in the genus *Aegilops* represent an important reservoir of genetic diversity for wheat improvement[5–7]. The genus comprises 23 species encompassing diploid C, D, M, N, S, and U genome lineages and several polyploid combinations[8,9]. *Aegilops peregrina* (Hack.) Maire & Weiller (syn. *Ae. variabilis* Eig) is an allotetraploid (2n=4x=28) with genome constitution $S^pS^pU^pU^p$, combining a U genome from *Ae. umbellulata* with S genome of Sitopsis origin[10]. This species is an important tertiary gene pool donor for wheat improvement. Genetic variation from *Ae. peregrina* has been used to introduce resistance to leaf rust[11,12], stripe rust[13,14], powdery mildew[15], cereal cyst nematode[16], and root knot nematode[17]. Notably examples included, leaf rust resistance genes *Lr59*, *LrP*, and *LrAp*, stripe rust resistance gene *YrP* and nematode resistance genes *CreX*, *CreY*, and *Rkn*[9]. *Ae. peregrina* has also been found to be tolerant to abiotic stresses such as drought[18] and salinity[19] and it has been reported to have increased grain protein and micronutrient content[9]. This species contains substantial resistance to highly virulent stem rust races[20]. *Ae. peregrina* accession PI 604178 was selected in this present study based on its seedling drought tolerance response (Fig. 1A) and resistance to stem rust races[20] including TTKSK and TRTTF.

Here, we present a subgenome-resolved, chromosome-scale reference genome assembly of *Ae. peregrina* accession PI 604178 generated using PacBio HiFi and Hi-C sequencing data. The 10.13 Gb assembly contains 10.01 Gb (98.81%) anchored to 14 chromosome-scale pseudomolecules representing $1S^p$-$7S^p$ and $1U^p$-$7U^p$ chromosomes (Fig. 1B and 1C). The assembly has a contig N50 of 25.84 Mb and scaffold N50 of 746.48 Mb (Fig. 2A; Table 1). High quality of this assembly was supported by a consensus QV of 74.61, 97.83% k-mer completeness, and 99.9% complete BUSCOs. The independent BUSCO analyses of S^p and U^p subgenomes recovered 99.3% and 99.5% complete orthologs, respectively (Fig. 2B). LTR Assembly Index (LAI) values of 20.43 and 18.79 for S^p and U^p, respectively, showed high continuity across repeat-rich regions in two subgenomes (Table 1). Repetitive sequences comprise 85.93% of the chromosome-anchored assembly, with Gyspy, Copia, and CATCA elements representing the predominant transposable element (TE) groups (Fig. 1B and 3A; Table 2). Structural gene annotations identified 59,910 high-

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

confidence (HC) and 21,746 low-confidence (LC) protein coding genes (Fig. 3B; Table 3). We additionally assembled complete chloroplast and mitochondrial genomes in *Ae. peregrina*. This chromosome-scale reference genome establishes a genomic framework for the genetic and evolutionary study of *Ae. peregrina*. The resource will facilitate sequence-level characterization of disease and pest resistance loci introgressed into wheat, development of subgenome specific markers for tracking *Ae. peregrina* chromatin in wheat backgrounds, comparative analyses of S^p and U^p subgenomes with their diploid relatives, and investigation of genome evolution following allopolyploidization. It also expands the genomic resources available for exploiting *Aegilops* diversity in wheat improvement.

Methods

Plant material and DNA extraction

We selected *Ae. peregrina* accession PI 604178 for whole genome assembly based on its seedling drought tolerance response (Fig. 1A) as per in-house developed protocol described in Gudi et al., 2026[21] and resistance to stem rust races[20], such as TTKSK and TRTTF. The seeds were procured from the United States Department of Agriculture-Germplasm Resources Information Network (USDA-GRIN). The accession PI 604178 was originally collected from Berekhya, Southern Coastal Plain, HaDarom, Israel (31.66666667, 34.63333333) with the sea elevation of 80 meters. The single seed decent method was used to increase seeds for this accession and to maintain genetic purity. Plants were grown in greenhouse under 16 hours (h) of daylight with 22°C day and 18°C night temperature. Young leaves collected from three-week old plants were immediately frozen in liquid nitrogen and stored at -80°C until DNA extraction. The high molecular weight (HMW) DNA was extracted using a CTAB-Qiagen Genomic-tip method[22]. We used nanodrop spectrophotometer and Qubit dsDNA HS assay to assess the quality and quantity of the extracted HMW DNA before library preparation.

PacBio HiFi sequencing and genome assembly

The HMW DNA was sheared to 17 kb fragments with Diagenode Megaruptor and SMRTbell Express Template Prep kit 2.0 was used for PacBio libraries[23]. The fragments larger than 10 kb were selected using Sage BluePippin size selection kit and were sequenced on the PacBio Revio platform using three 30 h SMRT cells. The Sequel II Sequencing kit 2.0, Sequencing Primer v5, and Sequel Binding kit 2.2 were used. The HiFi sequencing generated 217.62 Gb of

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

data with an average read length of 14,782 bp and a median read quality score of Q33 (~99.95% accuracy). Approximately 21.5x PacBio HiFi genome coverage was produced. The genome size and k-mer characteristics were estimated using GenomeScope2 (v2.0.1)[24]. The k-mer frequency distributions were generated from HiFi data using Meryl (v1.4.2)[25]. GenomeScope2 analyses were performed at k-mer lengths of 31 and 41. At k=31, GenomeScope2 estimated a genome size of approximately 9.97 Gb with unique sequence fraction of 23.4%, and a k-mer error rate of 0.0912%. Analysis at k=41 produced a genome size estimate of approximately 10.01 Gb, with 31.6% unique sequence and a k-mer error rate of 0.105% (Fig. 4A). Estimated heterozygosity was approximately 0.001% in both analyses. The k-mer spectrum contained a major peak at approximately 21x coverage.

We used hifiasm (v0.25.0-r726)[26] to assemble PacBio HiFi reads into contigs with default parameters for an inbred species. The primary assembly consisted of 2,419 contigs spanning 10.19 Gb, with an average contig length of 4.22 Mb, and a largest contig length of 111.49 Mb. The assembly had a contig N50 of 25.84 Mb with a contig L50 of 115, indicating high continuity prior to Hi-C scaffolding (Table 1). The primary contig assembly contained no gaps prior to scaffolding.

Organelle de-novo assembly

Chloroplast and mitochondrial genomes were assembled independently from PacBio HiFi reads using Oatk (v1.0)[27]. The parameters "-k 1001 -c 100 -m embryophyta_mito.fam -p embryophyta_pltd.fam" were used. The chloroplast and mitochondrial genomes were assembled as single circular genome, with the size of 136,919 bp and 449,543 bp, respectively.

Hi-C sequencing and genome scaffolding

We generated 1,206,141,041 chromosome conformation capture (Hi-C) read pairs for scaffolding of primary contig assembly. The leaf tissue from the same three-week old plant used for PacBio sequencing were collected and submitted to Phase Genomics (Settle, WA) for Hi-C library preparation. Three Hi-C libraries were prepared following the standard Phase Genomics proximity ligation protocol. The crosslinked chromatin was digested with four-enzyme cocktail (DpnII, DdeI, HinfI, and MseI), subjected to proximity ligation, and prepared as 2 x 150 bp paired-end sequencing libraries for Illumina sequencing. The raw Hi-C reads were aligned to the primary contig assembly following the Arima Mapping Pipeline (A160156_v03, Arima Genomics, Carlsbad, California, USA). The mapping showed a high alignment rate with 99.06% of read pairs

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

having both mates mapped to primary assembly. Duplicate alignments were identified using SAMBLASTER (v0.1.26)[28], which marked 657,857,611 of the mapped pairs as duplicates. After filtering and duplicate removal, 536,891,087 non-duplicate Hi-C read pairs were retained for scaffolding. We used a proximity guided scaffolder, YaHS (v1.2a.2)[29], for scaffolding of our primary assembly and found 28,928,874 intra-chromosomal and 507,962,213 inter-chromosomal contacts. The manual inspection and correction was performed using JuiceBox (v2.15)[30]. The inter-contig gaps within the scaffolds were padded with 100Ns. A final Hi-C contact matrix was generated at 250 kb resolution.

Assembly quality assessment

To screen the *Ae. peregrina* assembly for any contaminants, FCS-GX[31] with "fcs.py screen genome" and "fcs-gx" database with *Ae. peregrina* taxonomic identifier 130461 was used, while adapter screening was performed with "run_fcsadaptor.sh --euk". No contaminants or adapter sequences were reported in the screening check. The de-novo assembled organelles were mapped to the scaffolded assembly using minimap2 (v2.30-r1287)[32] to identify mitochondrial and chloroplast sequences. The analysis identified 1,408 organelle-derived scaffolds, which were filtered out from the assembly. Additionally, short contigs (<20 kb) were excluded from the final assembly. In total, 1,478 scaffolds totaling 66.45 Mb were removed from the final nuclear genome assembly. The final assembly contained 614 scaffolds spanning 10.13 Gb, of which 10.01 Gb (98.81%) was assigned to 14 chromosome-level pseudomolecules representing seven S^p and seven U^p chromosomes (Fig. 1B and Table 1). The remaining 120.44 Mb was retained as unplaced scaffolds. For naming and orienting the pseudomolecules, *Ae. longissima* accession AEG-6782-2 (v2)[33] (S genome) and *Ae. umbellulata* accession PI 554389[34] (U genome) assemblies were used. We performed a telomere search on all 14 chromosomes using tidk (v0.2.31)[35] with "clade" option set to "Poales". The chromosome-level genome assembly of *Ae. peregrina* PI 604178 was evaluated for accuracy, completeness, and contiguity, using tools such as QUAST (v5.2.0)[36], assembly-stats (v1.0.1), Merqury (v1.4.1)[25], BUSCO (v 6.0.0)[37], and LTR_retriever (v3.0.4)[38]. The pairwise dotplots were generated with chromeister (v1.5.a)[39]. The plots in this study are generated with ggplot2[40] package in R (4.4.0) and figure panels were arranged in Inkscape.

Structural and functional annotations

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

We used well curated ClariTeRep database[41] for transposable element (TE) annotations. The RepeatMasker (v4.1.5)[42] was run using "-xsmall -xm -engine rmblast" parameters and the output was processed with "clariTE.pl" script (https://github.com/jdaron/clari-te) to merge fragmented matches, reconstruct nested insertions, resolve element boundaries and assign TE copies according to the ClariTeRep hierarchical classification. TE abundance and sequence coverage were calculated for 14 chromosomes (Table 2).

The soft-masked genome was used for structural annotations of gene models using BRAKER3[43]. We used BRAKER3 in ETP mode and integrated transcript, protein-homology, and *ab initio* evidence for protein-coding gene annotations. The transcript evidence consisted of publicly available *Ae. peregrina* RNA-seq datasets SRR16550937-SRR16550944[citation]. The "viridiplantae" (https://bioinf.uni-greifswald.de/bioinf/partitioned_odb12/Viridiplantae.fa.gz) protein evidence were downloaded from OrthoDB (v12)[37]. The predicted protein coding genes were functionally annotated with InterProScan (v5.38-76.0)[44] and putative gene functions were assigned in the gff3 files, based on the domain conserveness and sequence similarity with curated Unisprot-Sprot protein database (https://www.uniprot.org/help/downloads). The gene identifiers follow the format "AeP.PI604178.r1.[chromosome]G[GeneID]. For instance, "AeP.PI604178.r1.1SpG0000100" denotes a gene located on chromosome $1S^p$ of the PI 604178 release 1 annotation. We categorized the gene models into high-confidence (HC) and low-confidence (LC) groups based on the sequence similarity with the Magnoliopsida UniProt TrEMBL dataset and the Triticeae repetitive element database (TREP). HC genes were defined as those sharing significant similarity (E-value < 1e-10) with TrEMBL proteins, demonstrating > 66% amino acid identity and < 25% difference in sequence length[45]. Those gene models that missed these homology thresholds, lacked external support, or matched TE sequences in TREP database (E-value < 1e-10), were classified as LC. The chromosome-anchored HC annotation comprised 29,784 genes on the S^p subgenome and 29,543 genes on the U^p subgenome. The gene architecture was highly similar between the two subgenomes (Table 3).

Data Records

The genome assembly for *Ae. peregrina* accession PI 604178 is deployed to GrainGenes Genome Browser (https://graingenes.org/GG3/genome_browser)[46] with BLAST functionality. We deposited raw PacBio HiFi sequencing data generated for *Ae. peregrina* accession PI 604178 in

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

NCBI Sequence Read Archive (SRA) with accession numbers SRR40310784, SRR40310785, and SRR40310786. The Hi-C sequencing data was submitted to SRA with SRR40310783 accession number. Genome has been deposited at DDBJ/ENA/GenBank under the accession JCCJTD000000000. The version described in this paper is version JCCJTD010000000. The reference genome assembly is accessible through GenBank under GCA_########.1as a part of BioProject PRJNA1517304. Datasets such as mitochondrial and chloroplast genomes, soft-masked genome assembly, repeat annotations, structural gene models, and functional gene annotations are available via figshare.

Technical validation

The Hi-C contact map showed strong and continuous intra-chromosomal interaction signals across all 14 pseudomolecule diagonals, supporting accurate chromosome-scale contiguity (Fig. 4B). FCS-GX screening detected no foreign sequence contamination or adapter sequences. Merqury analysis showed a consensus QV of 74.61 and k-mer completeness of 97.83% (Fig. 2A). The k-mer copy-number spectrum showed that most reliable read-derived k-mers were represented in the assembly, with the predominant assembly-specific peak corresponding to single-copy representation (Fig. 5A). Chromosome-specific QV ranged from 77 to 82, demonstrating consistently high consensus accuracy across both S^p and U^p chromosomes (Fig. 5B). Telomeric repeat arrays were identified at 4 of 14 chromosome ends in the S^p subgenome and 10 of 14 chromosome ends in the U^p subgenome (Fig. 1C). Finaly, LAI values of 20.43 for S^p and 18.79 for U^p further supported high continuity of repeat-rich regions. Whole genome pairwise comparisons showed strong chromosome-level collinearity of the S^p subgenome with *Ae. longissima* and the U^p subgenome with *Ae. umbellulata*, supporting the assignment, ordering, and orientation of the 14 *Ae. peregrina* chromosomes (Fig. 6). The quality of the structural gene annotation was independently evaluated using BUSCO analysis of the HC protein set, which recovered 97.8% complete BUSCOs, with 0.1% fragmented and 2.1% missing orthologs, supporting the high completeness of the predicted protein-coding gene space (Fig. 2B). In contrast, the LC protein set recovered only 0.9% complete BUSCOs, consistent with the enrichment of conserved protein-coding gene space within the HC annotation. The HC gene annotation contained 59,910 genes, distributed nearly equally between the S^p (29,784) and U^p (29,543) subgenomes. Gene structural features were also highly comparable between the two subgenomes, including gene and CDS

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

lengths, exon and intron lengths, and exon number per gene (Table 3), indicating consistent annotation quality across both subgenomes.

Usage Notes

To allow community access, the *Ae. peregrina* genome assembly and annotations have been submitted to public repositories. Interactive genome viewer for this assembly is available through GrainGenes genome browser. The direct BLAST functionality is also available at https://graingenes.org/blast/.

Data Availability

The raw sequencing data, genome assembly, and gene model annotations have been deposited in public repositories. The data generated for *Aegilops peregrina* accession PI 604178 is submitted to NCBI under BioProject accession number PRJNA1517304 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1517304/). The raw Hi-C and PacBio HiFi sequencing data is accessible through SRA accession numbers from SRR40310783 to SRR40310786 (https://identifiers.org/ncbi/insdc.sra:SRP729778). The genome assembly is made available on GenBank, figshare, and GrainGenes Browser. The supporting datasets including gene annotations and organelle genome assemblies are accessible via figshare (https://doi.org/10.6084/m9.figshare.33330000).

Code Availability

In this study, we used publicly available bioinformatics tools, and no custom scripts were developed. We used default software parameters, and detailed information on software names and versions is provided in the Methods section. All command-line procedures and analyses can be reproduced using the listed software and parameters.

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

References

1. Shiferaw, B. *et al.* Crops that feed the world 10. Past successes and future challenges to the role played by wheat in global food security. *Food Sec.* **5**, 291–317 (2013).
2. Wan, H. *et al.* Genetic Improvement and Application Practices of Synthetic Hexaploid Wheat. *Genes* **14**, 283 (2023).
3. Bhatta, M., Morgounov, A., Belamkar, V., Poland, J. & Baenziger, P. S. Unlocking the novel genetic diversity and population structure of synthetic Hexaploid wheat. *BMC Genomics* **19**, 591 (2018).
4. Cox, T. S. Deepening the Wheat Gene Pool. *Journal of Crop Production* **1**, 1–25 (1997).
5. Kishii, M. An Update of Recent Use of Aegilops Species in Wheat Breeding. *Front. Plant Sci.* **10**, (2019).
6. Schneider, A., Molnár, I. & Molnár-Láng, M. Utilisation of Aegilops (goatgrass) species to widen the genetic diversity of cultivated wheat. *Euphytica* **163**, 1–19 (2008).
7. Gaurav, K. *et al.* Population genomic analysis of Aegilops tauschii identifies targets for bread wheat improvement. *Nat Biotechnol* **40**, 422–431 (2022).
8. Kilian, B. *et al.* Aegilops. in *Wild Crop Relatives: Genomic and Breeding Resources* (ed. Kole, C.) 1–76 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011). doi:10.1007/978-3-642-14228-4_1.
9. Badaeva, E. D., González Franco, M. J., Razumova, O., Tereshchenko, N. A. & Divashuk, M. Perspectives on the utilization of Aegilops species containing the U genome in wheat breeding: a review. *Front. Plant Sci.* **16**, (2025).
10. Badaeva, E. D. *et al.* Genome differentiation in Aegilops. 4. Evolution of the U-genome cluster. *Plant Syst. Evol.* **246**, 45–76 (2004).

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

11. Marais, G. F., McCallum, B. & Marais, A. S. Wheat leaf rust resistance gene Lr59 derived from Aegilops peregrina. *Plant Breeding* **127**, 340–345 (2008).

12. Narang, D. *et al.* Discovery and characterisation of a new leaf rust resistance gene introgressed in wheat from wild wheat Aegilops peregrina. *Sci Rep* **10**, 7573 (2020).

13. Narang, D. *et al.* Fine mapping of Aegilops peregrina co-segregating leaf and stripe rust resistance genes to distal-most end of 5DS. *Theor Appl Genet* **132**, 1473–1485 (2019).

14. Liu, D. *et al.* Transfer of stripe rust resistance from Aegilops variabilis to bread wheat.

15. Spetsov, P., Mingeot, D., Jacquemin, J. M., Samardjieva, K. & Marinova, E. Transfer of powdery mildew resistance from Aegilops variabilis into bread wheat. *Euphytica* **93**, 49–54 (1997).

16. Coriton, O. *et al.* Assignment of Aegilops variabilis Eig chromosomes and translocations carrying resistance to nematodes in wheat. *Genome* **52**, 338–346 (2009).

17. Yu, M. Q., Jahier, J. & Person-Dedryver, F. Chromosomal location of a gene (Rkn-mnl) for resistance to the root-knot nematode transferred into wheat from Aegilops variabilis. *Plant Breeding* **114**, 358–360 (1995).

18. Liu, C., Yang, Z. & Hu, Y.-G. Drought resistance of wheat alien chromosome addition lines evaluated by membership function value based on multiple traits and drought resistance index of grain yield. *Field Crops Research* **179**, 103–112 (2015).

19. Farooq, S., Niazi, M. L. K., Iqbal, N. & Shah, T. M. Salt tolerance potential of wild resources of the tribe Triticeae. *Plant Soil* **119**, 255–260 (1989).

20. Olivera, P. D., Rouse, M. N. & Jin, Y. Identification of New Sources of Resistance to Wheat Stem Rust in Aegilops spp. in the Tertiary Genepool of Wheat. *Front. Plant Sci.* **9**, 1719 (2018).

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

21. Gudi, S., Singh, J., Faris, J., Gill, U. & Gupta, R. Identification of genetic factors governing drought tolerance in wild and domesticated tetraploid wheat. *BMC Genomics* https://doi.org/10.1186/s12864-026-13221-9 (2026) doi:10.1186/s12864-026-13221-9.

22. Vaillancourt, B. & Buell, C. R. High molecular weight DNA isolation method from diverse plant species for use with Oxford Nanopore sequencing. 783159 Preprint at https://doi.org/10.1101/783159 (2019).

23. Singh, J. *et al.* A reference-grade chromosome-level genome assembly of a historically important U.S. bread wheat cultivar Timstein. *Sci Data* https://doi.org/10.1038/s41597-026-07930-9 (2026) doi:10.1038/s41597-026-07930-9.

24. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* **11**, 1432 (2020).

25. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245 (2020).

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

27. Zhou, C. *et al.* Oatk: a de novo assembly tool for complex plant organelle genomes. *Genome Biol* **26**, 235 (2025).

28. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).

29. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, btac808 (2023).

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

30. Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**, 99–101 (2016).

31. Astashyn, A. *et al.* Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol* **25**, 60 (2024).

32. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

33. Avni, R. *et al.* Genome sequences of three Aegilops species of the section Sitopsis reveal phylogenetic relationships and provide resources for wheat improvement. *The Plant Journal* **110**, 179–192 (2022).

34. Singh, J. *et al.* Genomes of *Aegilops umbellulata* provide new insights into unique structural variations and genetic diversity in the U-genome for wheat improvement. *Plant Biotechnology Journal* **22**, 3505–3519 (2024).

35. Brown, M. R., Manuel Gonzalez de La Rosa, P. & Blaxter, M. tidk: a toolkit to rapidly identify telomeric repeats from genomic datasets. *Bioinformatics* **41**, btaf049 (2025).

36. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).

37. Tegenfeldt, F. *et al.* OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. *Nucleic Acids Res* **53**, D516–D522 (2025).

38. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res* **46**, e126 (2018).

39. Pérez-Wohlfeil, E., Diaz-del-Pino, S. & Trelles, O. Ultra-fast genome comparison for large-scale genomic experiments. *Sci Rep* **9**, 10274 (2019).

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

40. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer International Publishing, Cham, 2016). doi:10.1007/978-3-319-24277-4.

41. Daron, J. *et al.* Organization and evolution of transposable elements along the bread wheat chromosome 3B. *Genome Biology* **15**, 546 (2014).

42. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-3.0. (1996).

43. Gabriel, L. *et al.* BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* **34**, 769–777 (2024).

44. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

45. Singh, J., Gudi, S., Maughan, P. J., Gupta, R. & Gill, U. Improved reannotations of Aegilops umbellulata (PI 554389) genome and transcriptomics data provide candidates for leaf rust resistance for wheat improvement. *Sci Rep* **15**, 44930 (2025).

46. Yao, E. *et al.* GrainGenes: genetics, genomes, and pangenomes. *Genetics* **232**, iyaf270 (2026).

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Author Contributions

R.G. and J.S conceived and designed the study. J.S., S.G., and P.J.M. performed data analyses. J.S. prepared the first draft of the manuscript. J.S., S.G., P.J.M, U.G., and R.G. contributed to data generation, interpretation of results, and manuscript editing. R.G. procured the funding and supervised the project. All authors reviewed and approved the final manuscript.

Competing Interests

The authors share no competing interests.

Funding

This research was supported by USDA Agricultural Research Service project number 3060-21000-046-000D. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer. This work used resources of the Center for Computationally Assisted Science and Technology (CCAST) at North Dakota State University, which were made possible in part by NSF MRI Award No. 2019077.

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Tables:

Table 1: Assembly and quality statistics of *Aegilops peregrina* genome.

	Aegilops peregrina
Assembly length (bp)	10,132,462,779
Total contigs	2,419
Contig N50 (bp)	25,837,692
Contig L50	115
Chromosome-anchored assembly (bp)	10,012,022,159
Chromosome-anchored genome (%)	98.81
Chromosomes	14
Largest chromosome (bp)	933,292,056
Scaffold N50 (bp)	746,478,991
Scaffold N90 (bp)	534,546,439
Scaffold L50	6
Assembly L90	13
Total gaps	683
GC content (%)	46.92
BUSCO (%) (Whole, S^p, and U^p subgenomes)	99.9, 99.3, and 99.5
High confidence genes	59,910
Low confidence genes	21,746
Repetitive sequence (%)	85.93
k-mer completeness (%)	97.83
Quality value (QV)	74.61
LAI (S^p and U^p subgenomes)	20.43 and 18.79
Mitochondrial genome length (bp)	449,543
Chloroplast genome length (bp)	136,919

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Table 2: Summary of transposable element (TE) composition in the *Aegilops peregrina* genome assembly.

	Aegilops peregrina		
	Number of elements	**Length (bp)**	**Genome proportion (%)**
All TEs	3,211,306	8,603,114,080	85.93%
Class 1	1,731,462	6,970,994,209	69.63%
LTR retrotransposons			
Gypsy (RLG)	1,083,766	4,832,078,711	48.26%
Copia (RLC)	381,927	1,691,069,212	16.89%
Unclassified RT-LTRs (RLX)	127,897	328,807,738	3.28%
Non-LTR retrotransposons			
LINE (RIX)	132,865	118,182,367	1.18%
SINE (SIX)	5,007	856,181	0.01%
Class 2	1,074,976	1,480,451,758	14.79%
DNA transposons			
CACTA (DTC)	688,408	1,363,594,975	13.62%
Mutator (DTM)	92,486	45,947,755	0.46%
Harbinger (DTH)	40,216	17,756,826	0.18%
Mariner (DTT)	140,816	18,731,287	0.19%
Unclassified with TIRs (DTX)	99,904	25,747,785	0.26%
Unclassified Class 2 (DXX)	10,002	6,319,159	0.06%
hAT (DTA)	1,530	1,066,261	0.01%
Helitrons (DHH)	1,614	1,287,710	0.01%
Unclassified repeats (XXX)	404,868	151,668,113	1.51%

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Table 3: Summary statistics of high confidence gene models in *Aegilops peregrina* genome.

Features	**S^p subgenome**	**U^p subgenome**	**Whole assembly**
Total genes	29,784	29,543	59,910
Total transcripts	32,851	32,406	65,912
Total exons	153,804	150,798	307,203
Average gene length (bp)	3,027	2,946	2,985
Median gene length (bp)	1,867	1,869	1,863
Average CDS length (bp)	1,295	1,292	1,293
Median CDS length (bp)	1,107	1,107	1,107
Average exon length (bp)	276.6	277.6	277.5
Median exon length (bp)	141	141	141
Average intron length (bp)	502	491.3	497.5
Median intron length (bp)	132	132	132
Exons per gene	5.16	5.1	5.13
Monoexonic genes	8,556	8,825	17,637
Multiexonic genes	21,228	20,718	42,273
Exons per transcript	4.68	4.65	4.66
Monoexonic transcripts	9,576	9,885	19,756
Multiexonic transcripts	23,275	22,521	46,156

Note: Whole assembly values also include gene models located on unplaced scaffolds.

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

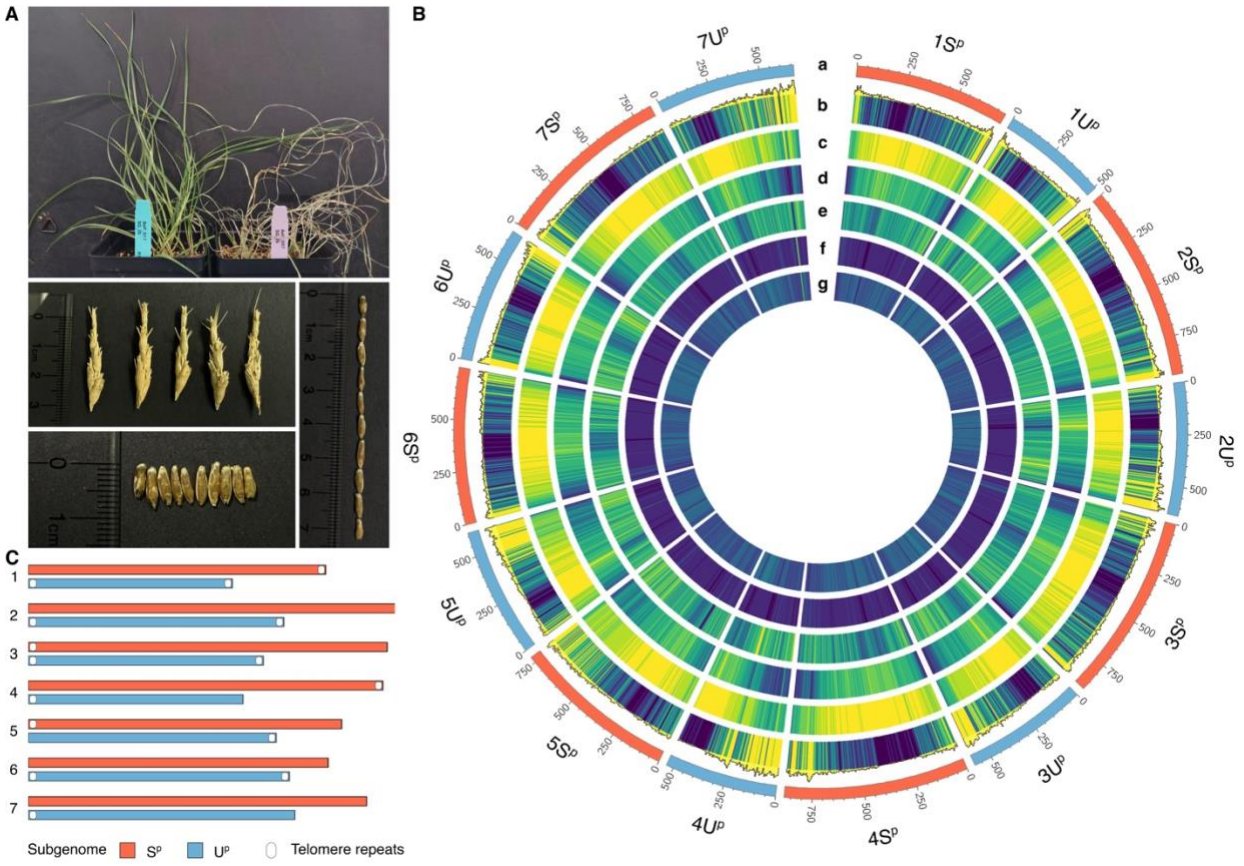


Fig. 1: Phenotype and chromosome-scale reference genome assembly of *Aegilops peregrina*. **(A)** *Ae. peregrina* accession PI 604178 (left) showing seedling drought tolerant response in comparison to susceptible accession (right), together with representative spikes and seeds of PI 604178. **(B)** Circular representation of 14 chromosomes comprising seven S^p and seven U^p subgenomes. Tracks from outside to inside represent: (a) chromosomes, (b) gene density, (c) repeat density, (d) Gypsy retrotransposon diversity, (e) Copia retrotransposon density, (f) CACTA density, and (g) GC content distribution. **(C)** Telomeric repeat arrays detected at the chromosome ends of S^p and U^p subgenomes.

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

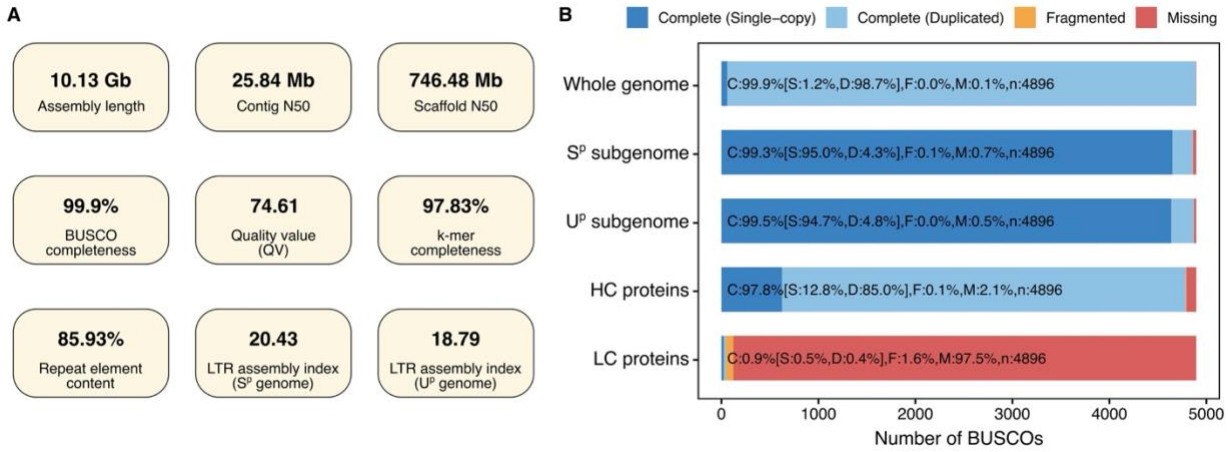


Fig. 2: Assembly and annotation quality assessment of *Ae. peregrina* reference genome. (A) Summary of major assembly quality metrics, including total assembly length, contig N50, scaffold N50, BUSCO completeness, consensus quality value (QV), k-mer completeness, repeat content, and LTR Assembly Index (LAI) for the Sᵖ and Uᵖ subgenomes. **(B)** BUSCO assessment using the poales_odb10 dataset (n = 4,896) for the whole-genome assembly, independently evaluated Sᵖ and Uᵖ subgenomes, and high-confidence (HC) and low-confidence (LC) predicted protein sets.

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

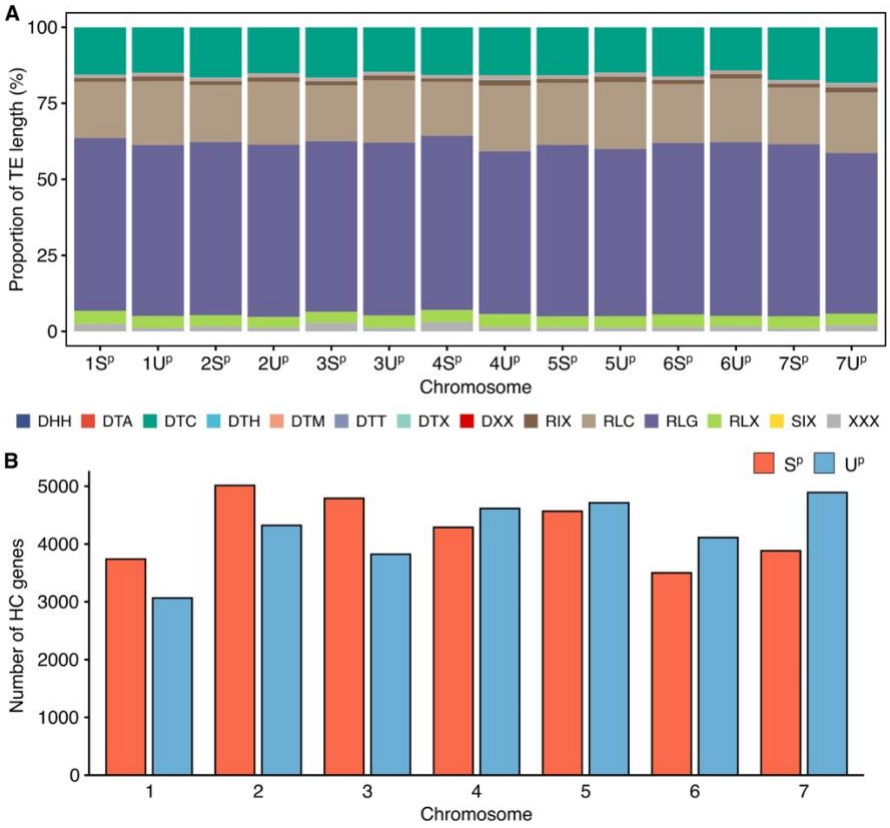


Fig. 3: Transposable element composition and high-confidence gene distribution across the *Ae. peregrina* chromosomes. (A) Relative contribution of major transposable element classes to the repeat-annotated sequence of each Sᵖ and Uᵖ chromosome. **(B)** Number of high-confidence (HC) protein-coding genes on corresponding Sᵖ and Uᵖ chromosomes for each of the seven homoeologous groups.

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

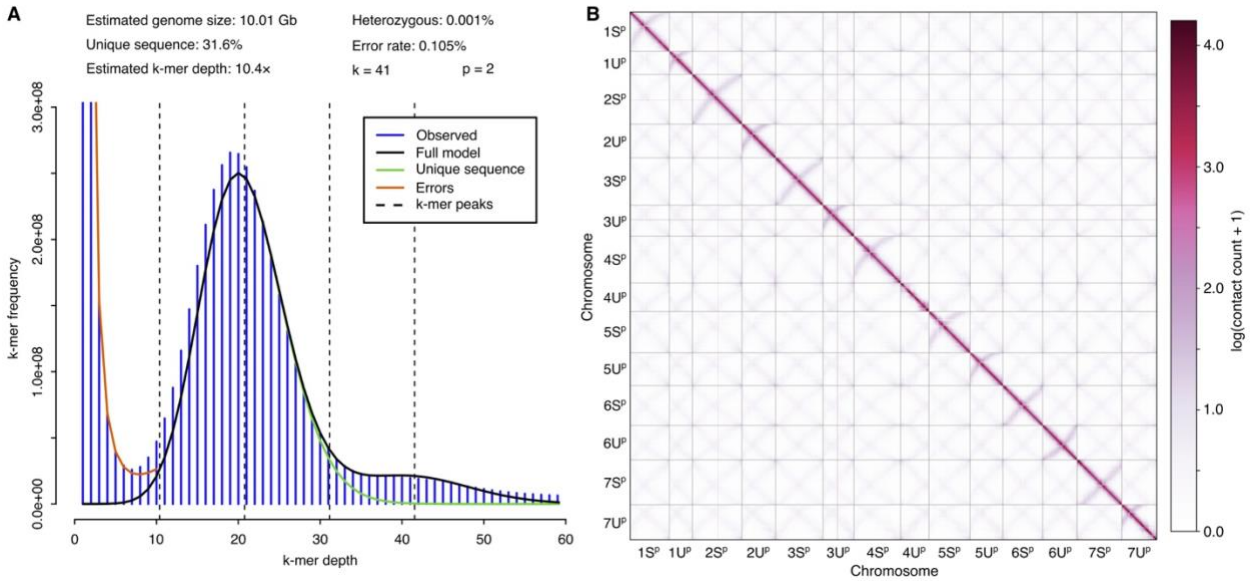


Fig. 4: Genome k-mer characteristics and Hi-C validation of chromosome-scale *Ae. peregrina* assembly. (A) GenomeScope2 k-mer frequency distribution generated from PacBio HiFi reads at k = 41. **(B)** Hi-C contact matrix of 14 chromosomes at 250 kb resolution.

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

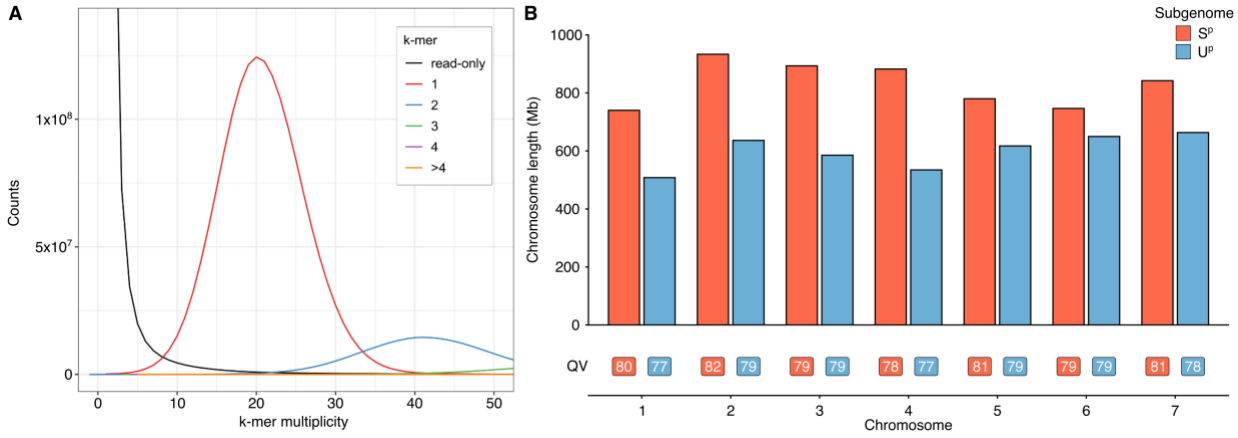


Fig. 5: k-mer based evaluation of assembly completeness and chromosome-level consensus accuracy. (A) Merqury k-mer copy-number spectrum comparing PacBio HiFi read-derived k-mers with their representation in the final assembly. Curves represent read-only k-mers and k-mers occurring one, two, three, four, or more than four times in the assembly. **(B)** Lengths and Merqury consensus quality values (QV) of corresponding S^p and U^p chromosomes across the seven homoeologous groups.

bioRxiv preprint doi: https://doi.org/10.64898/2026.08.28.747929; this version posted August 30, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

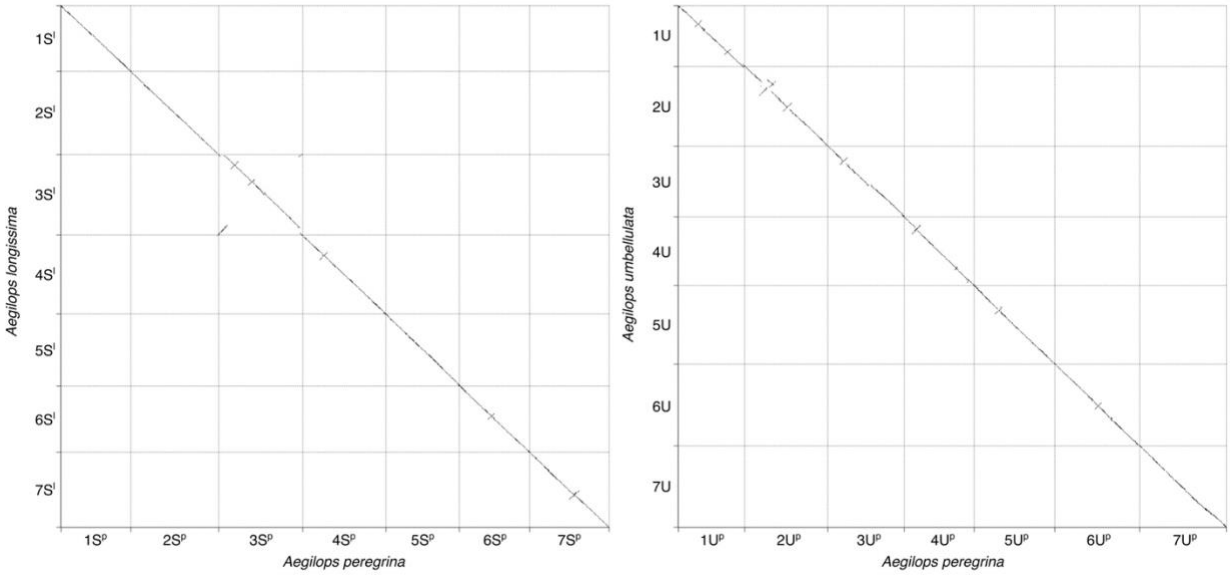


Fig. 6: Chromosome-level collinearity of *Ae. peregrina* S^p and U^p subgenomes with diploid relatives. Whole genome dotplot comparisons of the seven *Ae. peregrina* S^p chromosomes with the S genome chromosomes of *Ae. longissima* (left) and the seven U^p chromosomes with the U-genome chromosomes of *Ae. umbellulata* (right). Continuous chromosome-scale diagonals demonstrate extensive macro-collinearity and support the assignment, ordering, and orientation of the *Ae. peregrina* pseudomolecules.