Genomes of Aegilops umbellulata provide new insights into unique structural variations and genetic diversity in the U-genome for wheat improvement

Jatinder Singh, Santosh Gudi, Peter J Maughan, Zhaohui Liu, James Kolmer, Meinan Wang, Xianming Chen, Matthew N Rouse, Pauline Lasserre-zuber, Héléne Rimbert, et al.

HAL Id: hal-04707299
https://hal.inrae.fr/hal-04707299v1

Submitted on 24 Sep 2024

Genomes of *Aegilops umbellulata* provide new insights into unique structural variations and genetic diversity in the U-genome for wheat improvement

Jatinder Singh[1] (iD), Santosh Gudi[1] (iD), Peter J. Maughan[2], Zhaohui Liu[1], James Kolmer[3], Meinan Wang[4], Xianming Chen[4,5], Matthew N. Rouse[3,6], Pauline Lasserre-Zuber[7], Héléne Rimbert[7], Sunish Sehgal[8], Jason D. Fiedler[9] (iD), Frédéric Choulet[7], Maricelis Acevedo[10], Rajeev Gupta[1,9],* (iD) and Upinder Gill[1],* (iD)

[1]North Dakota State University, Fargo, North Dakota, USA

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

[3]Cereal Disease Laboratory, United States Department of Agriculture (USDA) Agricultural Research Service (ARS), St. Paul, Minnesota, USA

[4]Washington State University, Pullman, Washington, USA

[5]Wheat Health, Genetics, and Quality Research Unit, United States Department of Agriculture (USDA) Agricultural Research Service (ARS), Pullman, Washington, USA

[6]University of Minnesota, St. Paul, Minnesota, USA

[7]Université Clermont Auvergne, INRAE, GDEC, Clermont-Ferrand, France

[8]South Dakota State University, Brookings, South Dakota, USA

[9]United States Department of Agriculture (USDA) Agricultural Research Service (ARS), Cereal Crops Research Unit, Edward T. Schafer Agriculture Research Center, Fargo, North Dakota, USA

[10]Cornell University, Ithaca, New York, USA

Summary

Aegilops umbellulata serve as an important reservoir for novel biotic and abiotic stress tolerance for wheat improvement. However, chromosomal rearrangements and evolutionary trajectory of this species remain to be elucidated. Here, we present a comprehensive investigation into *Ae. umbellulata* genome by generating a high-quality near telomere-to-telomere genome assembly of PI 554389 and resequencing 20 additional *Ae. umbellulata* genomes representing diverse geographical and phenotypic variations. Our analysis unveils complex chromosomal rearrangements, most prominently in 4U and 6U chromosomes, delineating a distinct evolutionary trajectory of *Ae. umbellulata* from wheat and its relatives. Furthermore, our data rectified the erroneous naming of chromosomes 4U and 6U in the past and highlighted multiple major evolutionary events that led to the present-day U-genome. Resequencing of diverse *Ae. umbellulata* accessions revealed high genetic diversity within the species, partitioning into three distinct evolutionary sub-populations and supported by extensive phenotypic variability in resistance against several races/pathotypes of five major wheat diseases. Disease evaluations indicated the presence of several novel resistance genes in the resequenced lines for future studies. Resequencing also resulted in the identification of six new haplotypes for *Lr9*, the first resistance gene cloned from *Ae. umbellulata*. The extensive genomic and phenotypic resources presented in this study will expedite the future genetic exploration of *Ae. umbellulata*, facilitating efforts aimed at enhancing resiliency and productivity in wheat.

Introduction

Wheat wild relatives (WWRs) serve as an important source of stress resiliency for wheat improvement. *Aegilops umbellulata* Zhuk. is one such diploid (2n = 2x = 14) WWR with a large and repetitive U-genome similar to other *Triticeae* genomes (Li *et al.*, 2022; Zimin *et al.*, 2017). *Ae. umbellulata*, first reported by Zhukovsky (1928), is a self-pollinated annual grass species that grows primarily in the subtropical ecosphere. It is naturally distributed more prominently in Turkey but has also spread to other West Asian countries along the Fertile Crescent, including Iraq, Lebanon, Iran (West), Syria (North); and Caucasus region – Azerbaijan and Armenia (Eldarov *et al.*, 2015; Kilian *et al.*, 2011). Despite its narrow geographical distribution, *Ae. umbellulata* harbours substantial genetic diversity greater than *Ae. tauschii*, the D-genome donor of cultivated wheat (Okada *et al.*, 2018). *Ae. umbellulata* played a key role as a diploid progenitor in the evolution of seven polyploid species belonging to section *Pleionathera*, such as *Ae. triuncialis* L. (U^tC^t), *Ae. peregrina* Hack. (U^pS^p), *Ae. kotschyi* Boiss. (U^kS^k), *Ae. columnaris* Zhuk. (U^cX^c), *Ae. geniculata* Roth. (U^gM^g), *Ae. biuncialis* Vis. (U^bM^b), *Ae. neglecta* Req. ex Bertol. (U^nM^n), and *Ae. neglecta* (U^nM^nN^n) (Badaeva *et al.*, 2004).

Multiple genome analyses have been conducted to understand the evolution and dynamics of the U-genome in diploid and polyploid species (Badaeva, 2002; Badaeva et al., 2004; Kimber and Yen, 1989). For instance, Kimber and Yen (1989) explored the chromosomal pairing in interspecies hybrids between autotetraploid *Ae. umbellulata* (UUUU) and different natural polyploid species containing the U-genome. Similarly, Badaeva and co-authors (2004) employed heterochromatin banding patterns such as C-banding and fluorescence in situ hybridization to study the phylogenetic relationship between *Ae. umbellulata* and the seven polyploid species sharing the U-genome. These studies revealed that the U-genome in all tetraploid and hexaploid species is derived from *Ae. umbellulata* and is highly stable in all polyploid species. However, the other genomes (viz., C, S, M, or N) present in species belonging to *Pleionathera* section tend to be modified due to genome instability (Badaeva et al., 2004). These genomic instabilities might have occurred due to the meiotic instability induced during homoeologous chromosome pairing (SanMiguel et al., 1996) or due to increased activity of transposable elements (TEs) (Griffiths et al., 2006; Kerber, 1964; Riley and Chapman, 1958).

Due to high genetic diversity and resilience to multiple biotic and abiotic stresses, several efforts have been made to introgress agronomically important traits, including stress resilience and improved baking quality, from *Ae. umbellulata* into cultivated wheat (Bansal et al., 2017, 2020; Chhuneja et al., 2007; Edae et al., 2016; Edae and Rouse, 2019; Sears, 1956; Wang et al., 2018, 2022). In 1950s, first resistance gene, *Lr9*, was transferred from *Ae. umbellulata* to cultivated wheat through X-ray irradiation (Sears, 1956). Subsequently, other resistance genes from *Ae. umbellulata*, including *Lr76* (leaf rust), *Yr70* (stripe rust), and *PmY39* (powdery mildew) were successfully introgressed into wheat (Bansal et al., 2020; Zhu et al., 2006). Resistance to destructive stem rust races TTTTF and TTKSK (Ug99) was recently reported in *Ae. umbellulata* and a major quantitative trait locus was mapped on chromosome 2U (Edae et al., 2016). Besides disease resistance, allosyndetic pairing was induced between chromosome 1U of *Ae. umbellulata* and group 1 chromosomes of wheat to transfer high-molecular weight glutenin subunit (HMW-GS) genes (*1Ux* and *1Uy*) (Liu et al., 2002; Wang et al., 2018). Song et al. (2023), recently reported that, *Ae. umbellulata* has high zinc, iron, and seed gluten content compared to hexaploid wheat and *Ae. tauschii*. The same study also revealed higher levels of resistance to wheat stripe rust (Song et al., 2023). In addition to fungal pathogens, *Ae. umbellulata* was found to carry resistance to Hessian fly (Gill et al., 1985). Despite its huge potential to improve wheat, the genomic resources for *Ae. umbellulata* are not well established to effectively study and utilize novel traits/genes residing in this species except for recently published *Ae. umbellulata* genome in a parallel study (Abrouk et al., 2023).

Here we presented a comprehensive analysis of the structural variations and the evolution of U-genome by generating a high-quality chromosome scale assembly of *Ae. umbellulata* accession, 'PI 554389' that confers resistant to multiple wheat diseases. Comparative genome analysis across the *Triticeae* group revealed major chromosomal rearrangements unique to *Ae. umbellulata*. Furthermore, we resequenced 20 geographically and morphologically diverse *Ae. umbellulata* accessions with varying levels of resistance to multiple wheat diseases including, leaf rust, stripe rust, stem rust, tan spot, and bacterial leaf streak (BLS) to understand the genetic diversity in *Ae. umbellulata* species. Our genotypic and phenotypic evaluation revealed the presence of six allelic variants of *Lr9* (the only leaf rust resistance gene cloned from *Ae. umbellulata* to date), along with additional leaf rust resistance genes, in these accessions.

Results

Near telomere-to-telomere genome assembly and gene annotations of *Ae. umbellulata*

Ae. umbellulata accession, 'PI 554389' was used to generate a near telomere-to-telomere (T2T) chromosome level genome assembly. This accession was originally collected from Turkey in 1991 and was found to confer high levels of resistance to multiple wheat diseases including leaf rust, stem rust, and tan spot (Figure 1a). A total of 4.28 Gb of *Ae. umbellulata* genome was assembled using: (i) PacBio HiFi reads (126.7 Gb) generated by circular consensus sequencing (CCS); (ii) Nanopore long reads (12.81 Gb) generated by Oxford Nanopore MinION sequencing; and (iii) High-throughput chromosome conformation capture (Hi-C) derived short-reads (Figure S1a; Tables S1–S3). Initially, a primary contig assembly was generated using the hifiasm assembler and HiFi data only. The primary contig assembly consisted of 1389 contigs with an N50 and L50 of 8.79 Mb and 141 respectively. Hi-C scaffolding of the primary contig assembly yielded seven pseudomolecules covering 4.20 Gb with N50 value of 634.10 Mb (Table 1; Table S4). The assembly was further polished by closing 252 gaps present in seven pseudomolecules using PacBio and ultra-long Nanopore reads (Table S5). The pseudomolecules were oriented and assigned chromosome names based on their homology with the IWGSC RefSeq_v2.1 D sub-genome (The International Wheat Genome Sequencing Consortium [IWGSC], 2018) (Figure 1b; Table S6). Of the total seven chromosomes, we found telomeres on both ends for four chromosomes (4U, 5U, 6U, and 7U) and one end for remaining three chromosomes (1U, 2U, and 3U) (Figure S1b). Chromosome sizes ranged from 488.85 Mb (1U) to 658.84 Mb (6U) and the GC content ranged from 46.91% (7U) to 47.3% (6U) (Figure 1b; Tables S7 and S8). Quality assessment of the assembled nuclear genome using INSPECTOR revealed a 99.99% mapping rate of HiFi reads with genome quality score of 59.54. Genome

Figure 1 Genome sequencing of *Aegilops umbellulata* accession, PI 554389: (a) phenotypic appearance and the disease reaction of *Ae. umbellulata* accession, PI 554389 to leaf rust and tan spot disease. TNBJS, MNPSD, TDBJQ and TBBGS are the leaf rust races affecting *Triticum aestivum*, BBBQD is the leaf rust races affecting durum wheat, and Pti2 is the race causing tan spot in wheat; (b) Circos plot showing the genomic features of *Ae. umbellulata* genome. Within the circos plot: **circle-i** depicts the individual *Ae. umbellulata* chromosomes with their physical positions (Mb), short arm (light grey), centromere (red) and long arm (dark grey); **circle-ii** depicts the number (histogram) and density (heatmap) of high-confidence genes; **circle-iii** depicts the percentage distribution of transposable elements (TEs); **circle-iv** depicts the percentage distribution of Gypsy (RLG) elements; **circle-v** depicts the percentage distribution of Copia (RLC) elements; **circle-vi** depicts the percentage distribution of CACTA (DTC) elements; **circle-vii** depicts the percentage of GC content; and **circle-viii** depicts the number of resistant gene analogs (RGA).
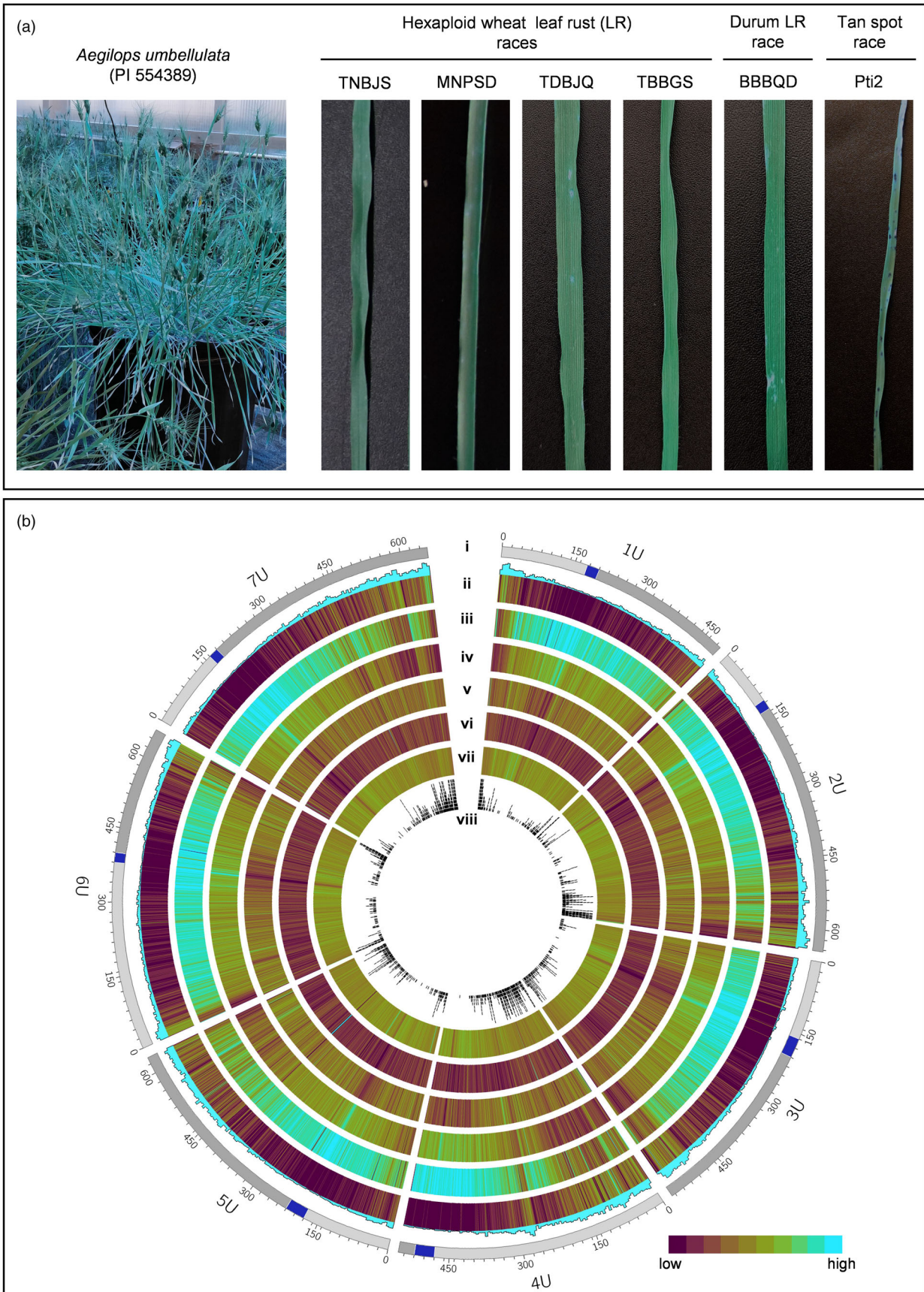
Table 1 Statistics of genome features of *Ae. umbellulata* chromosome level assembly

Category	Features	Values
Genome	Total genome length (Gb)	4.28
	Total length of chromosomes (Gb)	4.2
	GC content (%)	47.15
	N50 length (Mb)	634.1
Repetitive sequence	Total transposable elements (%)	84.1
	Retrotransposons (%)	69.1
	DNA transposons (%)	13.9
Protein-coding genes	Predicted protein-coding genes	78 076
	High confidence genes	48 366
	Low confidence genes	29 710

completeness was assessed using BUSCO, which identified that >98% of single copy conserved orthologs within the embryophyta dataset were identified as complete, with less than 5% being duplicated, as expected for a diploid species (Figure S1c; Table S9).

The full-length transcript data were generated using PacBio iso-seq single molecule real-time (SMRT) sequencing technology by pooling RNA samples from four different tissues (seedling shoots, seedling roots, leaves after vernalizations, and immature spikes) and used as primary evidence to annotate 78 076 gene models, of which, 48 366 were high-confidence (HC) genes (Table 1; Figure 1b; Table S7). Among the HC genes, a total of 2162 were predicted to be resistance gene analogs (RGAs), including 790 nucleotide-binding domain leucine-rich repeats, 1094 receptor-like kinases, 160 transmembrane-coiled-coil, and 117 receptor-like proteins (Li *et al.*, 2016) (Table S10).

Unique transposon landscape shapes the evolutionary trajectory of *Ae. umbellulata* species

TEs constitute major portions of eukaryotic genomes and play a significant role in genome evolution. In the PI 554389 genome, we identified 1.24 million copies of TEs which are present on non-orthologous positions in comparison to A, B, and D genomes of *T. aestivum* and account for 3.6 Gb (84.1%) of the whole U-genome (Figure S2; Table S11). Despite the near-complete TE turnover, the proportion of each superfamily is similar to that already observed in A-B-D sub-genomes with Gypsy, Copia, and CACTAs representing 47.1%, 17.7%, and 12.7% of total TEs respectively (Figure S2; Table S11). At the family level, only 5% (15/300 families) are enriched/underrepresented in the U-genome compared to A, B or D: five LTR-retrotransposons, one LINE, two DNA transposons, and seven unclassified repeats. These are low copy-number families, and we did not observe any massive burst of a particular family. Overall, 95% of the families have a copy number that is similar among U-A-B-D lineages, although not present at orthologous sites. The TE turnover was not accompanied by lineage-specific TE burst/loss. In general, TE families tend to maintain a relatively constant copy number and the U-genome is another example of *Triticeae* that followed this equilibrium model of evolution.

Comparative genome analysis revealed major chromosomal rearrangements unique to the U-genome

Comparative genome analysis is a powerful tool to dissect the genomic architecture and the evolutionary history of a species. Rooted species tree analysis using multi-copy orthologous genes among 11 *Triticeae* species and one outgroup species revealed three clades belonging to A-, B- and D-lineages, where *Ae. umbellulata* was phylogenetically placed in D-lineage (Figure 2a; Table S12). Among the D-lineage species, *Ae. umbellulata* (U-genome) diverged earlier followed by the separation of the D- and S-genome species from a common ancestral parent. During this evolutionary separation, the U-genome accumulated major chromosomal rearrangements (Zhang *et al.*, 1998) which led to broader genetic diversity in the U-genome (Okada *et al.*, 2018), as evident from the morphological and genetic variations in *Ae. umbellulata*. To further understand these structural variations, we performed a whole genome comparative analysis of the U-genome within the D-lineage species and outside of the D-lineage, including *Ae. speltoides*, *T. urartu*, A-B-D sub-genomes of wheat, and phylogenetically distant species *H. vulgare* and *B. distachyon*. We observed high levels of chromosomal discordance among studied species relative to the U-genome (Figure 2b; Figure S3). To study these structural variations and evolutionary relationship of the U-genome with A-, B- and D-lineages, we performed the macro-synteny analysis of *Ae. umbellulata* with *T. urartu*, *Ae. spletoides*, *Ae. tauschii*, *Ae. sharonensis* and *Ae. longissima*. Our results revealed major chromosomal rearrangements unique to the U-genome including several prominent non-reciprocal translocations in 4U and 6U chromosomes (Figure 2c; Figure S4). Based on synteny analysis among *Ae. tauschii*, *Ae. umbellulata*, and *Ae. longissima*, we identified a unique 7S^l/4S^l translocation (Figure 2c). This translocation has been reported previously only in *Ae. longissima* but absent in the rest of the S and D genome species (Li *et al.*, 2022). We also observed 5U/6U reciprocal translocation in *Ae. umbellulata* compared to other species except for *T. urartu* (Figure S4). In addition to translocations, two major intra-chromosomal inversions on 7U and 2U and multiple inter-chromosomal translocated inversions were detected (Figure 2c). In contrast, groups 1, 3, and 5 remained remarkably conserved across the U-, A-, B-, D- and S-genomes. All these structural variations have led to the formation of a unique U-genome of the *Ae. umbellulata* compared to other *Triticeae* species (Okada *et al.*, 2018; Sasanuma *et al.*, 2004).

Complex chromosomal rearrangements drive the evolution of present day 4U and 6U chromosomes

We observed pronounced chromosomal rearrangements particularly in chromosomes 4U and 6U. Therefore, chromosomes 4U and 6U were further examined in detail for their collinearity with two D lineage species, *Ae. tauschii* (D) and *Ae. longissima* (S^l) (Figure 3a; Figure S5). The previously known 7S^l/4S^l translocation in *Ae. longissima*, prompted us to study the S^l genome along with the D genome of *Ae. tauschii* (Ankori and Zohary, 1962; Li *et al.*, 2022; Ruban and Badaeva, 2018). Our results suggest the presence of ancestral chromosomal segments of 1L (long arm), 2L, 7L and 6S (short arm) on 4U and segments of 5L and 4L on 6U (Figure 3a; Figure S5). Based on these findings, we developed a model to explain the evolution of the 4U and 6U chromosomes (Figure 3b, c). The collinearity analysis suggests that a significant portion of ancestral chromosome 6S (~120 Mb) may have fused to chromosome 4S. In the subsequent non-reciprocal, inter-chromosomal translocation events, distal ends of ancestral chromosomes 7L (~55 Mb), 2L (~34 Mb) and 1L (~55 Mb) inversely fused to chromosome 4 of *Ae. umbellulata*, which gave birth to the modern day 4U chromosome of *Ae. umbellulata* (Figure 3b). Similarly, 6U chromosome may have evolved through non-reciprocal translocation of a substantial region of the
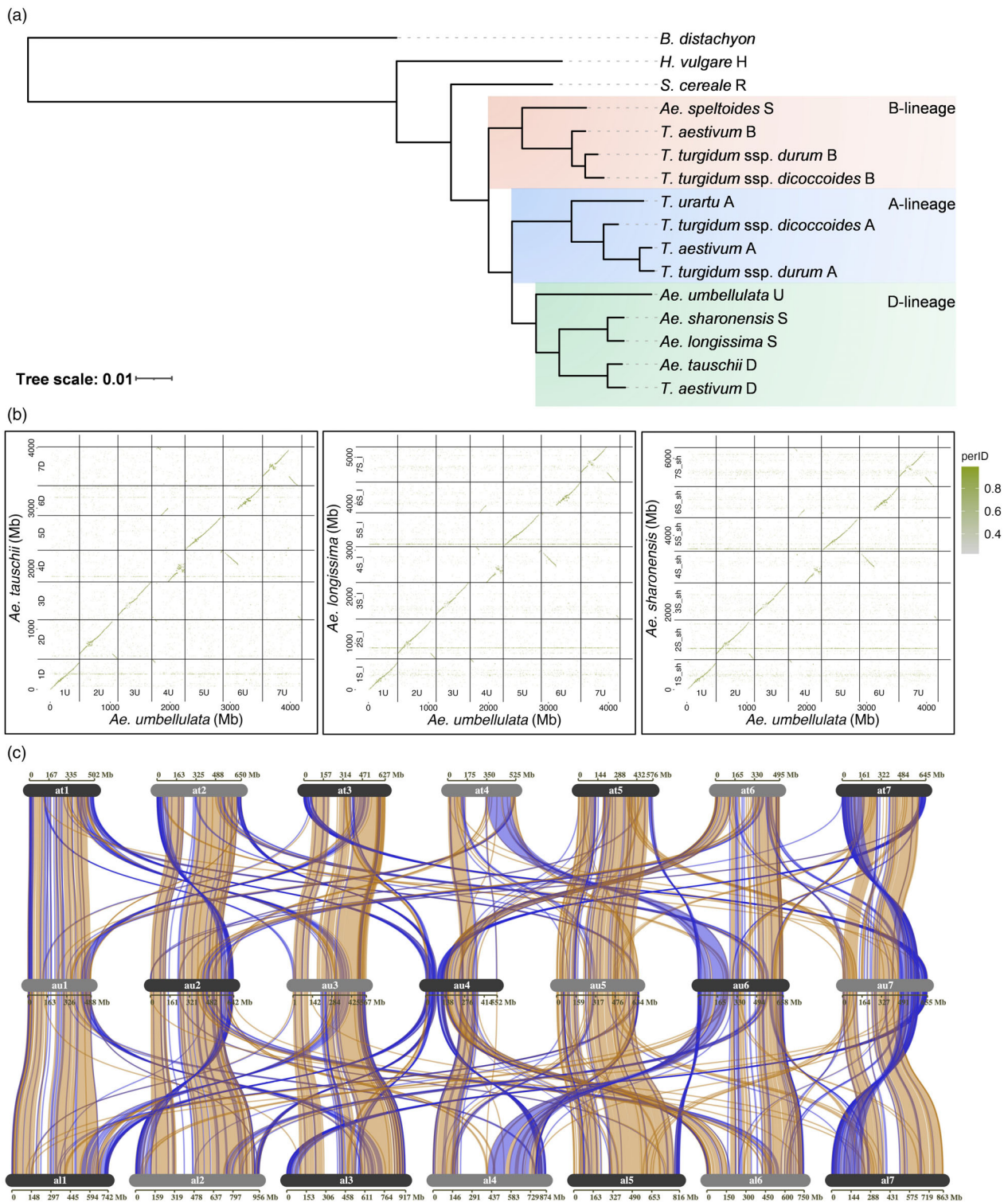
Figure 2 Comparative genome analysis of *Aegilops umbellulata* with other species: (a) dendrogram showing the phylogenetic relationship of *Ae. umbellulata* with 10 *Triticae* spp. and one outgroup species *Brachypodium distachyon*; (b) whole genome comparison of *Ae. umbellulata* with *Ae. tauschii*, *Ae. longissima* and *Ae. sharonensis*; (c) synteny and collinearity analysis of *Ae. umbellulata* (au) with *Ae. tauschii* (at) and *Ae. longissima* (al).

ancestral 4L chromosome (~210 Mb) into broken/wounded chromosome 6 resulted from translocation of 6S segment to chromosome 4 (Figure 3c). Later on, a part of translocated 4L segment on the 6U chromosome underwent a reciprocal translocation with the 5U chromosome, exchanging a region of

~35 Mb. These events suggest that 4U and 6U have evolved via independent sequential translocation events rather than a single major translocation event. In addition to these major translocation events, there were multiple intra-chromosomal rearrangements during the evolution of 4U and 6U chromosomes.
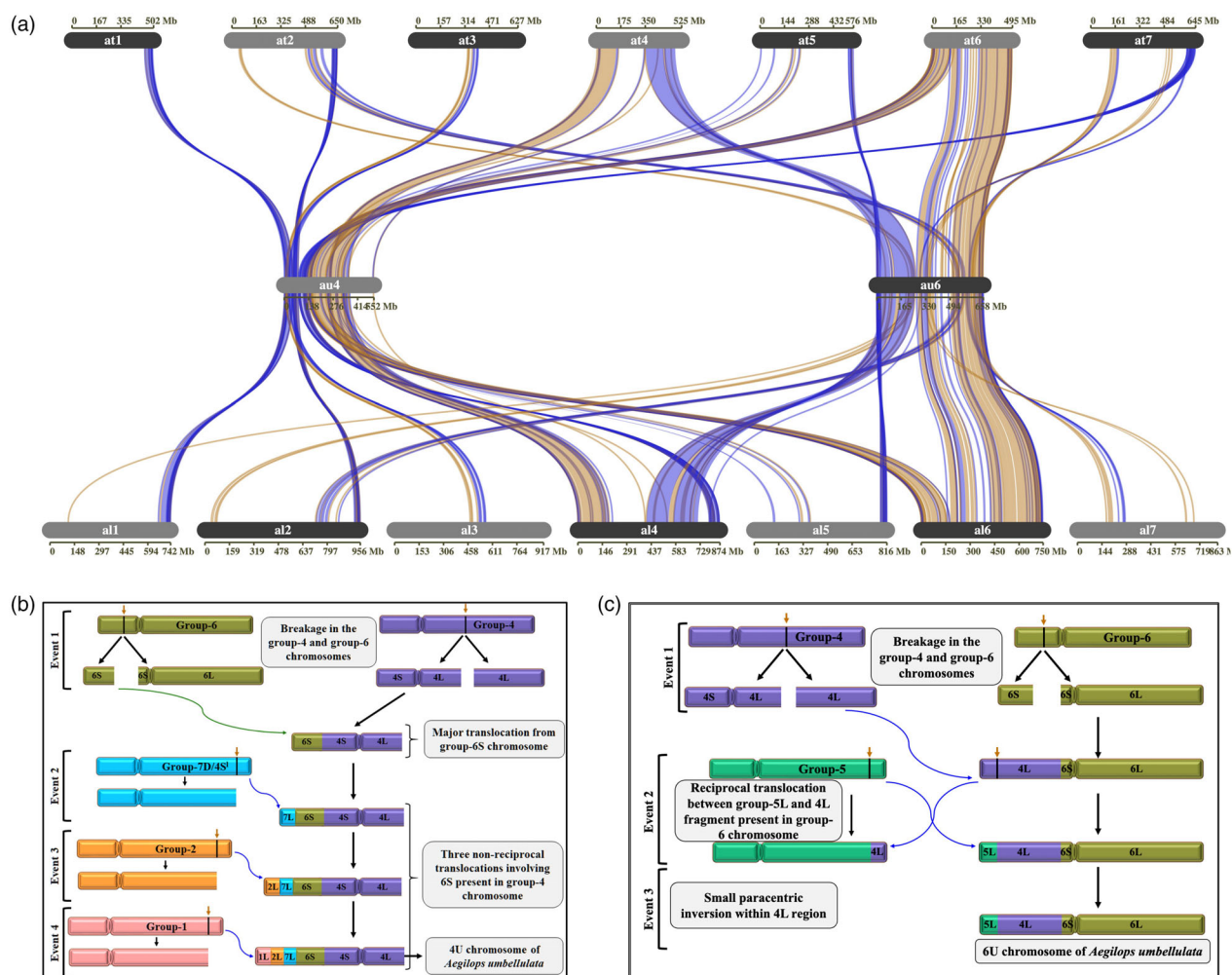
Figure 3 Evolution of 4U and 6U chromosomes of *Aegilops umbellulata*: (a) synteny and collinearity analysis of 4U and 6U chromosomes of *Ae. umbellulata* (au) with all chromosomes of *Ae. tauschii* (at) and *Ae. longissima* (al); (b) model depicting the origin of the 4U chromosome of *Ae. umbellulata*; (c) model depicting the origin of the 6U chromosome of *Ae. umbellulata*.

Resequencing of *Ae. umbellulata* highlights rich genetic diversity in the U-genome

To investigate the genetic diversity in *Ae. umbellulata*, 20 accessions representing a large geographical distribution and displaying high variability for morphology and resistance to five wheat diseases – leaf rust, stripe rust, stem rust, tan spot, and BLS were selected for whole genome resequencing (Figure 4a, b; Table S13; Figures S6–S9). Genome sequencing at ~10X coverage followed by variant calling using GATK (v.4.1.8.0) identified a total 86 931 487 SNPs (Table S14). After hard masking and filtering SNPs for read depth (4 < DP >15), SNP cluster (>3 SNPs within 10 bp window), non-biallelic SNPs, and SNPs unanchored to chromosomes with missing value of 90%, a total of 7 184 562 SNPs were retained and used for further analyses (Table S14). Principal component analysis (PCA) using the filtered sets of SNPs divided the 20 accessions into three sub-groups (Figure 4c). Fifteen accessions were present in group-I, whereas only three and two accessions were present in the group-II and group-III respectively. Most of the accessions collected from Turkey and all accessions from other regions fell in the group-I, except for the accessions collected from Serbia (group-II) and the United

Kingdom (group-III). Accessions in the group-I have comparatively higher level of resistance to various leaf rust races, whereas the accessions in group-II and III were susceptible to several leaf rust races (Figure 4c; Table S13). PCA results were further confirmed by estimating the maximum likelihood method for genotype grouping and ancestry coefficients using cross-entropy values, which revealed the presence of three sub-groups ($K = 3$) (Figure 4d, f). Nucleotide diversity (?) analysis among 20 accessions as well as within each sub-group revealed the maximum genetic diversity in group-III (? = 0.00035) and minimum genetic diversity in group-I (? = 0.0001) (Figure 4g). Additionally, the pairwise nucleotide diversity (?) analysis revealed variable nucleotide diversity within 1 Mb sliding window on each chromosome and as expected the diversity is relatively low around the centromeric regions and high near the chromosome ends (Figure S10).

Resequencing and disease evaluations revealed new allelic variants of historically important *Lr9* along with novel leaf rust resistance genes

Phenotypic and genotypic evaluations were performed to determine the presence of *Lr9* and additional leaf rust resistance

genes in the 20 *Ae. umbellulata* accessions. PI 554389 and seven additional accessions were highly resistant to 10 leaf rust races including the two races (MNPSD and TNBJS) that are virulent on *Lr9* (Figure 4a; Table S13.), suggesting the presence of *Lr9* and/or new leaf rust resistance gene(s). Two accessions, PI 554282 and PI 542375, were resistant to all tested leaf rust races except TNBJS and TBBGS, respectively, suggesting the presence of two or more novel *Lr* genes. Only one accession, 48114, had a reaction pattern similar to *Lr9*, while the remaining accessions displayed five unique reaction patterns indicating the presence of *Lr9* and additional *Lr* genes in *Ae. umbellulata* that are yet to be identified.

Since the *Lr9* gene has been cloned (Wang *et al.*, 2023), we searched for *Lr9* alleles in the 20 resequenced accessions. A complete copy of *Lr9* was present in 10 accessions including PI 554389 (Table S15). Based on the amino acid sequence variations from the cloned *Lr9* gene, we categorized the *Lr9* alleles into six haplotypes (Figure 5; Table S15). Remaining accessions, which lacked the complete *Lr9* coverage. The haplotype analysis of *Lr9* revealed the high conserveness in protein kinase I (PK-I) domain in all six haplotypes. In contrast, several amino acid substitutions were detected in protein kinase 2 (PK-II), Von Willebrand factor type A (vWA) and vWaint domains (Figure 5; Tables S15 and S16; Figure S11). A single amino acid substitution from Leucine to Proline (L642P) in the PK-II domain was found in all six haplotypes compared to the cloned *Lr9* (Figure 5; Table S16; Figure S11). It is possible that a natural mutation had occurred in the original source (TA1851) of *Lr9* before it was introgressed into the hexaploid wheat cultivar 'Thatcher'. The L642P substitution also defines haplotype-I (Hap-I) and is present in three accessions, 48114, 48646 and PI 554395. The phenotypic virulence pattern of 10 leaf rust races also supports the presence of a functional *Lr9* allele in 48114 due to its susceptibility only to *Lr9* virulent races, MNPSD and TNBJS (Table S13). Therefore, despite the single amino acid substitution, Hap-I is a functional allele of *Lr9*. In contrast, the other two accessions, 48646 and PI 554395, harbouring Hap-I are resistant to all tested leaf rust races, including the *Lr9* virulent races, MNPSD and TNBJS, suggesting the presence of a novel leaf rust gene(s) along with *Lr9* in these accessions. Hap-II, represented by PI 542369, has a single amino acid insertion of threonine (S21_S22insT) along with the aforementioned L642P substitution (Figure 5; Table S16; Figure S11). PI 542369 displayed a leaf rust reaction similar to 48114 except for an additional susceptibility to MHDSB which suggests that Hap-II may also be a functional allele of *Lr9* but with minor alterations in the virulence pattern. Hap-III, Hap-IV, Hap-V and Hap-VI have amino acid substitutions mainly in PK-II or PK-II and vWA or PK-II, vWA and vWaint domains. Accessions harbouring the Hap-III (PI 554389), Hap-IV (110746), Hap-V (PI 298905) and Hap-VI (PI 573515) were completely resistant to all the tested leaf rust races including the races (MNPSD and TNBJS) virulent on *Lr9*. It is possible that these lines may have functional alleles of *Lr9* along with an additional unknown *Lr* gene(s). We also identified two accessions (PI 554417 and AE 1595) which had the complete *Lr9* gene, with premature stop codon before the PK-I domain (Table S15). PI 554417 was highly susceptible to all tested races and AE 1595 was moderate in response to four out of 10 races including a race virulent on *Lr9*, suggesting the absence of a functional *Lr9* allele in both accessions (Tables S13 and S15).

Discussion

The genus, *Aegilops* has been extensively explored to harness useful traits for bread wheat improvements (Schneider *et al.*, 2008). *Ae. umbellulata* is one of the WWRs that carries large diversity for stress resiliency. At the chromosomal level, *Ae. umbellulata* has structural rearrangements that are unique to this species (Song *et al.*, 2020; Zhang *et al.*, 1998). To study these changes and the associated genetic diversity, a T2T genome assembly and a single-base resolution genomic variation map of 20 resequenced *Ae. umbellulata* accessions were developed in this study. In *Ae. umbellulata*, 84.1% of the genome was made up of repetitive elements, which is very similar to the previous findings for A, B and D sub-genomes of *T. aestivum* and genomes from other related tetraploid and diploid *Triticeae* species with TEs varying from 83 to 88% (Figure S2; Table S11) (Papon *et al.*, 2023). The TEs in U-genome are not conserved at the orthologous positions of A, B and D genomes. Lack of conserved positions of TEs were also observed among A, B and D genomes due to the TE turnover, that is, cycle of deletions/amplifications during ~5 Myrs of divergence, as explained earlier (Wicker *et al.*, 2018). Higher proportion of repetitive elements in *Triticeae* are major drivers that have shaped the genomic landscape of wheat and WWRs (Bariah *et al.*, 2020; Wicker *et al.*, 2022). In the present study, we found a unique TEs landscape in *Ae. umbellulata*, which has little conserved orthology with wheat TEs. This difference in the distribution of *Ae. umbellulata* TEs is likely to have led to considerable chromosomal structural rearrangements that characterize the U-genome. TEs such as RLG_Sabrina, RLG_WHAM and RLC_Angela are already known to induce structural and functional variations in rye and wheat (Rabanus-Wallace *et al.*, 2021; Wicker *et al.*, 2022). TE-mediated rearrangements are also well studied in human genomes and play an important role in the species evolution (Balachandran *et al.*, 2022). Phylogenetically, U-genome is grouped in D-lineage, but formed a distinct clade (Figure 2). A previously known 7S^l/4S^l translocation in *Ae. longissima* (Ruban and Badaeva, 2018) was also observed in *Ae. umbellulata* although it appeared to be a distinct and independent event. The rearrangements in chromosome 4U and 6U are especially significant, where major segments of 4U and 6U chromosomes have experienced extensive reciprocal and non-reciprocal translocation events (Figure 3). Based on the collinearity analysis with *Ae. tauschii* and *Ae. longissima*, the present day 4U and 6U chromosomes appear massively restructured, consisting of large portions from multiple chromosomes, specifically 1L-2L-7L-6S-4S-centromere-4L and 5L-4L-6S-centromere-6L respectively (Figure 3). The centromeric regions of chromosome 4U and 6U are syntenic to wheat group-4 and 6 chromosomes respectively. In the absence of a high-quality genome, previous studies were unable to correctly identify the origin of these chromosomes, resulting in erroneously interchanging of chromosome names (Abrouk *et al.*, 2023; Athwal and Kimber, 1972). In the present study, we rectified this error and assigned the correct names to these chromosomes to facilitate future genomics studies in *Triticeae*. This correction is also supported by evidences presented in the previous studies (Yang *et al.*, 1996; Zhang *et al.*, 1998).

The complex chromosomal rearrangements observed in *Ae. umbellulata* might have played a role in the genetic diversity observed in the species even though it is distributed across a narrow geographical area (Badaeva *et al.*, 2004; Molnár *et al.*,
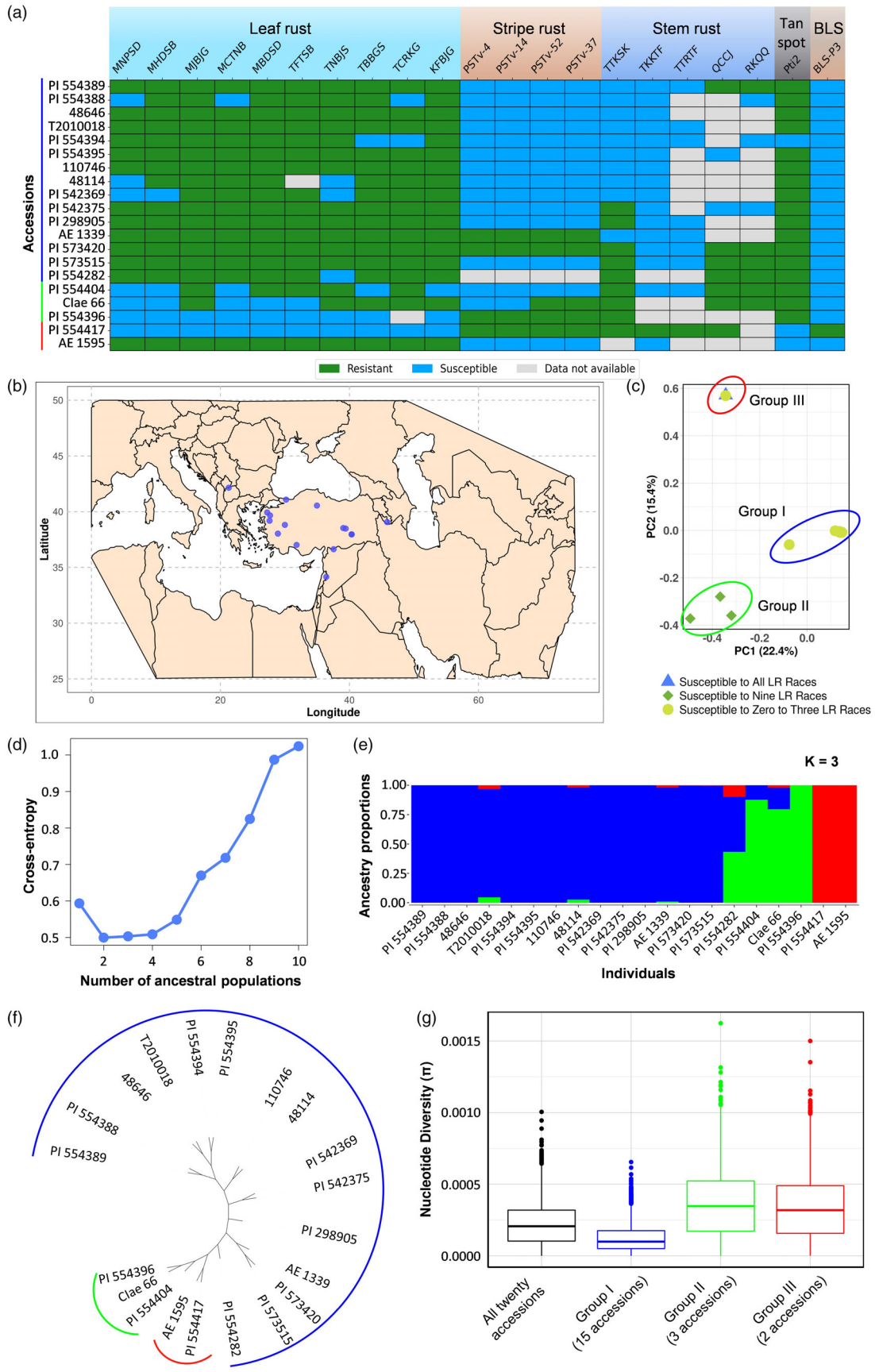
Figure 4 Assessing genetic diversity among 20 *Aegilops umbellulata* accessions: (a) disease reaction of 20 *Ae. umbellulata* accessions for multiple races of leaf rust (10 races), stripe rust (4 races), stem rust (5 races), tan spot (1 race) and bacterial leaf streak (BLS); (b) geographical distribution of 20 *Ae. umbellulata* accessions; (c) principal component analysis (PCA) to study the population stratification among *Ae. umbellulata* accessions. PCA divided all accessions into three subgroups with 15 (Group I), 3 (Group II) and 2 (Group III) accessions in each subgroup; (d) cross entropy values showing the number of ancestral populations among the *Ae. umbellulata* accessions; (e) landscape and ecological association (LEA) analysis for studying the population diversity among *Ae. umbellulata* accessions. Where, K is the number of subgroups in the population; (f) maximum-likelihood method for grouping *Ae. umbellulata* accessions into subgroups; (g) nucleotide diversity (?) analysis to estimate the amount of genetic diversity residing among all *Ae. umbellulata* accessions as well as within each subgroup.
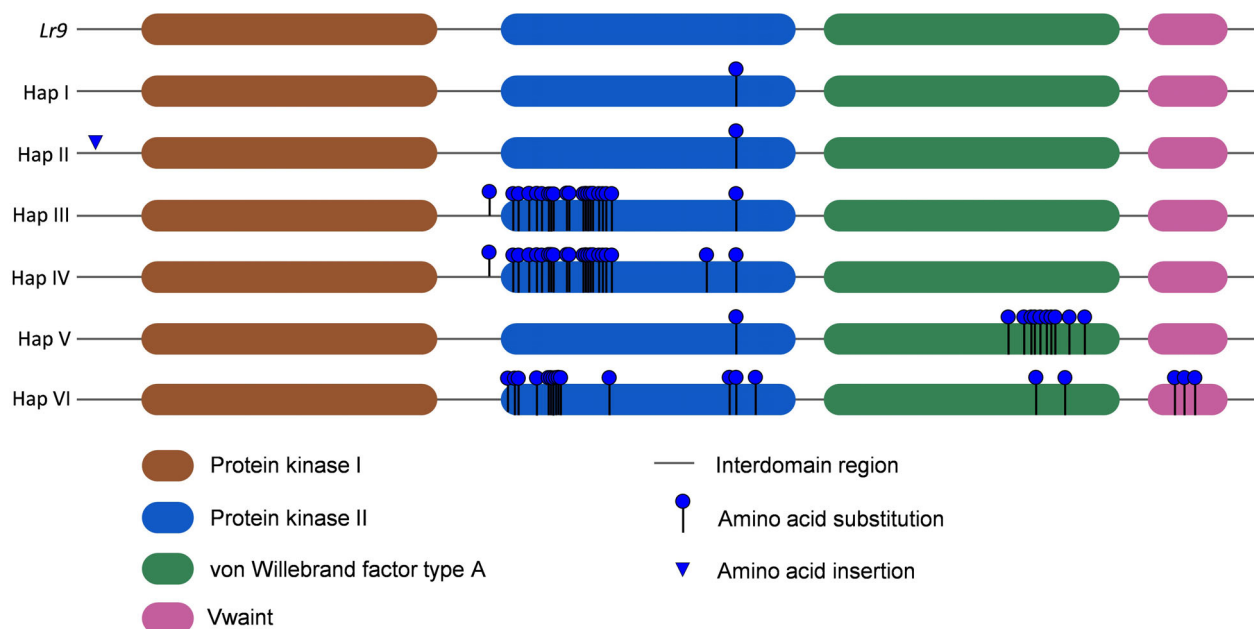
Figure 5 Protein sequence comparison of six LR9 haplotypes (Hap) identified from 20 *Aegilops umbellulata* accessions. Solid bars indicate conserved domains, inverted pyramid indicates indel, and red pinheads represent single amino acid substitutions in comparison to the cloned LR9 protein.

2016; Okada *et al.*, 2018). Higher genetic diversity was also reported in *Ae. umbellulata* compared to *Ae. tauschii* which was reflected by significantly higher levels of observed genetic polymorphism as evidenced by increased numbers of identified SNP and the occurrence of alleles with rare frequencies (Okada *et al.*, 2018; Sasanuma *et al.*, 2004). This increase in genetic diversity has also increased the variation observed in plant morphological features such as spike length, spike shape, plant growth, etc. and resistance response against different wheat diseases, *viz.* tan spot, BLS and three wheat rusts (Figure 4a; Figures S6–S9). These traits can be harnessed to widen the genetic base of modern cultivated bread wheat. Previous efforts in this direction have led to identification and introgression of several economically important genes including *Lr9* from *Ae. umbellulata* to bread wheat. Our phenotypic data on a small set of accessions supports the presence of at least five novel leaf rust resistance genes and stem rust resistance against highly virulent Ug99 and non-Ug99 lineages of stem rust belonging to different *Pgt* clades I, III and IV (Guo *et al.*, 2022; Olivera *et al.*, 2015, 2019; Szabo *et al.*, 2022). One out of 20 accessions (PI 554417) displayed high level of resistance to BLS which threatens the wheat production in the US and other parts of the world (Friskop *et al.*, 2023). Some level of genetic resistance to BLS is found in triticale, but most of the tested

wheat germplasm were shown to be susceptible (Sapkota *et al.*, 2018).

In a recent study, Wang *et al.* (2023) reported the cloning and characterization of a leaf rust resistance gene, *Lr9*. This gene was originally introgressed to wheat on chromosome 6BL from *Ae. umbellulata* by Sears in the 1950s and was reported to be located on chromosome 6U (Sears, 1956; Wang *et al.*, 2023). The resulting introgression on chromosome 6BL and aforementioned erroneous naming of chromosome 4U and 6U, might led to the confusion about the chromosomal position of *Lr9*. In the present study, we reported that 4U contains chromosomal segments from chromosomes 6, 4, 2 and 7, and *Lr9* is located between 72439902-72455974 bp on 4U in a segment translocated from the long arm of group-2 chromosome. In five radiation-induced introgression lines (carrying *Lr9*) originally developed by Sears, the *Lr9* segments were reported on 6B in two lines and on 4B, 2D and 7B in each line of wheat (Friebe *et al.*, 1995). Therefore, we speculate that the introgression of *Lr9* is a homoeologous exchange of chromosomal segments instead of random incorpo-ration (Zhang *et al.*, 1998). We located *Lr9* at chromosome 4U at 72439902–72455974 bp, which is homoeologous to the long arm of the wheat group-2 chromosomes. In our study, we have also identified six new haplotypes of *Lr9* with amino acid substitutions mainly in the PK-II, vWA, and Vwaint domains.

Among these domains, PK-II is assumed to be a pseudo-kinase (Wang et al., 2023). The functionality of these haplotypes can be evaluated in future studies.

In summary, our study has unveiled major evolutionary changes in U-genome over millions of years, which have resulted in enriched genetic diversity in this species. The genomic resources developed in this study present future opportunities to harness novel traits to build wheat cultivars with elevated resilience to biotic and abiotic stresses. The complex chromosomal rearrangements in *Ae. umbellulata* revealed in this study open new avenues to investigate the evolution of wheat tertiary gene pool species.

Materials and methods

Plant material and disease resistance assessment

Aegilops umbellulata accession 'PI 554389' was selected for whole genome sequencing based on the resistance responses to 10 hexaploid leaf rust races, one durum leaf rust race, two stem rust races and one tan spot race (Figure 1a; Table S13). The accession was procured from the United States Department of Agriculture-Germplasm Resources Information Network (USDA-GRIN), which was originally collected near the Hazar Lake, 35 km South East of Elazığ, Turkey. Disease evaluations of 20 *Ae. umbellulata* were performed under greenhouse conditions at seedling stages. Details of disease phenotypes are provided in Methods S1.

Genome assembly

PacBio HiFi library preparation and DNA sequencing

Aegilops umbellulata seedling was grown in a controlled growth chamber adjusted with 12 h (h) photoperiod and day/night temperature set to 20°/18 °C. The hydroponic growth solution was made using MaxiBloom® Hydroponics Plant Food (General Hydroponics, Sevastopol, CA, United States) at a concentration of 1.7 g/L. In preparation for PacBio HiFi sequencing, high-molecular weight (HMW) DNA was extracted from 2-week-old young leaf tissues given 72-h dark-treatment using a CTAB-Qiagen Genomic-tip protocol as described previously (Vaillancourt and Buell, 2019). DNA quantity and quality was checked using Qubit dsDNA HS assay and Nanodrop spectrophotometer respectively. HMW genomic DNA was sheared to 17 kb on a Diagenode Megaruptor and then made into SMRTbell adapted libraries using SMRTbell Express Template Prep Kit 2.0. Size selection was performed using a Sage BluePippin to select fragments greater than 10 kb and then sequenced at the BYU DNA Sequencing Center (Provo, UT, USA) using Sequel II Sequencing Kit 2.0 with Sequencing Primer v5 and Sequel Binding kit 2.2 for 30 h with adaptive loading using PacBio SMRT Link recommendations.

Further, Oxford Nanopore sequencing was used to generate super long reads on MinION sequencer using three FLO-MIN112 flow cells. HMW DNA was used for library preparation using SQK-LSK112 ligation kit for nanopore sequencing. The flow cell was primed according to the manufacturer's instructions and the library was loaded on the flow cell in a dropwise manner (Table S2). The raw data were basecalled using Guppy (v6.0.1; Oxford Nanopore Technologies, UK).

Genome assembly and scaffolding

A primary contig assembly was constructed using the PacBio HiFi CCS reads using Hifiasm v16.1 (Cheng et al., 2022) with the default parameters for an inbred species (−l 0). Hi-C was used to scaffold the primary contigs assembly into pseudo-molecules. Hi-C reads were aligned to the primary contig assembly using the Burrows–Wheeler Aligner (Li and Durbin, 2010). Only paired end reads that were uniquely aligned to contigs were retained for downstream analyses. Contigs were clustered, ordered, and oriented using Proximo™, an adapted proximity-guided assembly platform based on the LACHESIS method with proprietary parameters developed at Phase Genomics (Bickhart et al., 2017; Burton et al., 2013; Peichel et al., 2016). Gaps between scaffolds within the scaffolded assembly were filled with 100 Ns.

Manual gap closing and chromosome assignment

We used long reads generated by PacBio HiFi sequencing and Nanopore MinION sequencing to close the gaps in pseudomolecules, after adapter trimming and filtering the low-quality reads. Gap closing was performed using command line BLASTN. The PacBio HiFi and long nanopore reads that spanned the entire gap were extracted based on their similarity to the 2 kb region on both sides of a gap. These extracted reads spanning the gap were visualized using Geneious Prime 2023.0.4 (https://www.geneious.com) for the confirmation. Then, these sequences were used to fill the gaps manually in the assembly. Custom script with more detail on gap closing is provided in the code availability statement. The gap closed pseudomolecules were assigned with chromosome names based on their synteny to *T. aestivum* subgenome D (IWGSC RefSeq_v2.1) using minimap2 (Li, 2018). For identifying telomeric repeats (TTAGGG/CCCTAA), we used FindTelomeres (https://github.com/JanaSperschneider/FindTelomeres) using default parameters.

Annotations of gene models and transposable elements

Iso-seq and transcriptome assembly

For transcriptome assembly, tissue samples were collected at different growth stages from seedlings, roots, leaf tissues after vernalization and immature spike in liquid nitrogen. The frozen samples were ground with a mortar and pestle. About 100 µg frozen and ground tissue was used to extract RNA using Qiagen RNeasy Plant Mini Kit (74904) following the manufacturer's instructions. The quantity and quality of extracted RNA first tested using Nanodrop spectrophotometer, which was further evaluated with Bioanalyzer. After quality check, RNA from all the samples was pooled in equal molar ratios to synthesize full length complementary DNA (cDNA) using a NEBNext® single cell/low input cDNA synthesis and amplification kit (E6421L) which uses a template switching method to generate full length cDNAs (New England BioLabs, Ipswich, MA, USA). IsoSeq libraries were prepared from the cDNA according to standard protocols using the SMRTbell v3.0 library prep kit (Menlo Park, CA, USA) and sequenced on a single SMRT cell 8 M using a PacBio Sequel II the DNA sequencing center at Brigham Young University (Provo, Utah, USA).

Genome annotation

Species specific repeats identified by RepeatModeler v1.0.11 and RepeatMasker v4.0.9 (Smit et al., 1996) was used to identify, classify and mask repeat elements within the assembled genome relative to the Repbase-derived RepeatMasker libraries (Dfam 3.0; 20 190 227; http://www.girinst.org/). The MAKER3 pipeline (Cantarel et al., 2008) was used to annotate the final genome assembly using as primary evidence for the *A. umbellulata*

transcriptome (see above) and as alternative evidence for the predicted gene models and their translated protein sequences from barley (Hv_Morex.pgsb.Jul2020; https://wheat.pw.usda.gov/GG3/content/morex-v3-files-2021), wheat (https://phytozome-next.jgi.doe.gov/info/Taestivumcv_ChineseSpring_v2_1), and panBrachypodium (https://phytozome-next.jgi.doe.gov/info/BdistachyonPangenome_v1) as well as all proteins in the highly curated Uniprot Swiss-Prot database (*n* = 561 176). *Ab initio* gene prediction was based on *A. umbellulata* specific and *O. sativa* hidden Markov models for Augustus and SNAP respectively. Putative gene function was identified using BLAST searches of the predicted peptide sequences against the Swiss-Prot database using MARKER's default cut-off values (1E−6). Gene models were also classified as high confidence and low confidence based on predicted proteins using custom python scripts. High confidence models exhibited (1) a blastp hit (<1.0E-10) against the Magnoliopsida TrEMBL database (https://www.uniprot.org/) (UniProt Consortium, 2021), (2) a query coverage and length within 25% of the subject coverage and length, and (3) >66% identity between the query and subject.

Annotation of TEs and comparison of TE family proportions between the U- and A-B-D sub-genomes

TEs were annotated using CLARITE (Daron *et al.*, 2014) with the same parameters as described initially for *T. aestivum* Chinese Spring (Wicker *et al.*, 2018) and other *Triticum/Aegilops* (Aury *et al.*, 2022; Papon *et al.*, 2023). Briefly, TEs were predicted by similarity-search against the ClariTeRep library (which includes TREP http://botserv2.uzh.ch/kelldata/trep-db/) using RepeatMasker (Smit *et al.*, 1996) and were then modelled with CLARITE that resolve overlapping predictions, merge adjacent fragments of a single element, and reconstruct patterns of nested insertions. Abundance of the different TE families of the U-genome was calculated by cumulating the length (in bps) of all predictions assigned to the same (sub)family by CLARITE. To investigate the specificity of the U-genome TE landscape compared to A, B and D sub-genomes, we calculated fold-changes of proportions for each family compared to that annotated by the same approach in Chinese Spring RefSeq v2.1a (Wicker *et al.*, 2018). Abundance of each family was represented by the percentage of bps relatively to the (sub)genome size (all scaffolds for *Ae. umbellulata* genome assembly; only pseudomolecules for *T. aestivum*, that is, not considering chrUn). Enrichment of TE families in the U-genome were investigated by computing the log2 ratios between proportions observed in U *versus* A, B or D. We considered a substantial change in abundance when log2 fold-change was <−2 or >2. Only families accounting for more than 100 kb in at least one of the compared (sub)genomes were considered (i.e. 300 out of 501 families).

Genome, transcriptome and annotation quality assessment

Inspector v1.0.2 (Chen *et al.*, 2021) was used to evaluate the quality of the Hi-C scaffolded assembly using the PacBio HiFi reads. Inspector efficiently identifies and can correct both small-scale miss assemblies (base correction) as well as large-scale structural assembly errors. Furthermore, the completeness of the corrected Hi-C assembly, transcriptome and annotation was assessed relative to conserved orthologous genes within Poaceae (poales_odb10) using BUSCO v5.0.0 (Manni *et al.*, 2021) with the 'long' argument which applies Augustus optimization for self-training (Stanke and Morgenstern, 2005).

Whole genome RGA prediction

To predict RGAs in the assembled genome, we used the docker image of RGAugury (Li *et al.*, 2016) (https://bitbucket.org/yaanlpc/rgaugury/src/master/) with the default parameters and databases such as pfam and interproscan.

Whole genome comparison with wheat and wild wheat relatives

Phylogenetic analysis

We used 10 species (*T. aestivum*, *T. turgidum* ssp. *dicoccoides*, *T. turgidum* ssp. *durum*, *T. urartu*, *Ae. tauschii*, *Ae. speltoides*, *Ae. longissima*, *Ae. sharonensis*, *Hordeum vulgare* and *Secale cereale*) along with an outgroup species *B. distachyon* to perform phylogenetic analysis to find phylogenetic topologies for newly sequenced *Ae. umbellulata*. We used Orthofinder (Emms and Kelly, 2019) (v2.5.5) to find orthologs from high confidence genes for phylogenetic analysis. The high confidence genes from all the species were filtered using the same criteria that was used for classifying *Ae. umbellulata* gene models. The default parameters were used for Orthofinder to locate the multi-locus orthologous genes and generate species tree. The phylogenetic tree was visualized using online tree visualization tool iTol (Letunic and Bork, 2021) (v6.8.1) (https://itol.embl.de/).

Whole genome analysis and collinearity and synteny analysis

For whole genome analysis, we used minimap2 (Li, 2018) (v2.24) to generate pairwise alignment between *Ae. umbellulata* and other genomes. A custom R script was used to plot the dot plots using R package ggplot2. Further, synteny analysis was performed using MCScanX (Wang *et al.*, 2012) (https://github.com/wyp1125/MCScanX). Blastp was used to generate the all vs all protein alignments and a custom python script was used to prepare bed files from gff files for each species. An online tool SynVisio (Bandi and Gutwin, 2020) (https://synvisio.github.io/) was used to visualize the MCScanX output and to generate synteny plots.

Whole genome sequencing (WGS) of the 20 *Ae. Umbellulata* accessions

Plant material and Illumina short-read sequencing

We selected 20 geographically diverse *Ae. umbellulata* accessions based on the disease phenotyping scores for five wheat diseases (Table S13) for whole genome sequencing. Around 400 mg of leaf tissue was collected from a 5-week-old plant in liquid nitrogen. Each accession was grown from a single seed to maintain genetic purity. The collected leaf samples were sent to Novogene for DNA extraction, library preparation and sequencing (150 bp paired-end) at 10x coverage using the Illumina HiSeq 4000.

Mapping and short variant calling

Reads were filtered and trimmed with bbduk from the BBTools software suite (JGI, v 39.01) to remove adapter sequences and retain read pairs with an average quality greater than 13 and length greater than 75. Reads were quality checked with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Filtered read pairs were aligned to the *Ae. umbellulata* with bowtie2 (https://doi.org/10.1038%2Fnmeth.1923) with the '-very sensitive-local' parameter. After filtering out the secondary and other low quality alignments (MapQ < 20), the BAM files were sorted according to genomic coordinates and a CSI index was generated for each accession with SAMtools (Danecek *et al.*, 2021)

(v1.15). Genome Analysis Toolkit (McKenna *et al.*, 2010) (GATK v4.1.8.1) (https://gatk.broadinstitute.org/hc/en-us) and Picard (v2.22.8) were used for SNP calling. First, index and sequence dictionary for reference *Ae. umbellulata* genome was generated using SAMtools 'faidx' command and GATK CreateSequenceDictionary tool respectively. Duplicate reads were marked with MarkDuplicatesSpark command and CSI index for each duplicate marked bam file was generated with SAMtools. These duplicated marked bam files for each accession were used for SNP calling individually using GATK HaplotypeCaller with -ERC GVCF option. GATK CombineGVCFs tool was used to combine all 20 variant called files for joint genotyping using GenotypeGVCFs tool. SNPs were pulled out from joint genotyping files using GATK SelectVariants tools. The SNPs were filtered using VariantFiltration tool with parameters (QD < 2.0, FS > 60.0, MQ < 40.00, MQRankSum < −12.5, ReadPosRankSum < −8.0, SOR >3.0) Furthermore, SNPs clusters, where 3 SNPs appeared in 10 bp window, were also filtered. VCFtools (Danecek *et al.*, 2011) (v0.1.14) was used to remove the filtered SNPs based on the filter tags, low and high average SNP depth ($4 \leq DP \geq 15$) and we kept only b-allelic SNPs sites, and SNPs on seven chromosomes (removed unanchored chromosomes). Further, we kept SNPs with 10% or less missingness using BCFtools (Danecek *et al.*, 2021) (v1.15).

Diversity analysis and population structure

For diversity analysis, 7 184 562 filtered SNPs were used for PCA using PLINK (Purcell *et al.*, 2007) (v1.9) with default parameters. For PCA plot, R (v4.0.3) package ggplot2 was used and colour coded according to the susceptibility response to the total number of leaf rust races out of 11 tested races (susceptible to 0–3 races, susceptible to nine races, susceptible to 11 races). The individual ancestry proportions were computed with the 'snmf' function of R package LEA (Frichot and François, 2015) using the entropy option. For each K value from 1 to 10 (where K is the count of ancestral populations), 10 independent runs were performed on all filtered SNPs. For phylogenetic analysis VCF file was converted to PHYLIP format using vcf2phylip (v2.8) python script (https://github.com/edgardomortiz/vcf2phylip). Further, RAxML (Stamatakis, 2014) (v8.2.12) was used to generate 500 bootstrap trees (−f a -# 500 -m GTRCAT) and maximum likelihood tree (-D). Final phylogenetic tree was constructed by merging those trees (-f b -z -t -m GTRCAT) and visualized using iTOL (Letunic and Bork, 2021) (v6.8.1) (https://itol.embl.de/). To compute the nucleotide diversity (pi) for all 20 accessions, group I (15 accessions), group II (3 accessions) and group III (2 accessions), we used 'window-pi' option of VCFtools (Danecek *et al.*, 2011) (v0.1.14) over 1 000 000 bp sliding window and used R package ggplot2 for plotting nucleotide diversity. Further, to check pairwise nucleotide diversity between reference genome accession and other 19 accessions, we used the same above-mentioned VCFtools parameters and for chromosome wise plotting of each comparison, python3 libraries pandas and matplotlib were used. To locate the geographical location of *Ae. umbellulata* accessions on a map, we plotted the available latitude and longitude data of 17 accessions (Table S13) using R package ggplot2, maps and mapdata.

Lr9 exon capturing using WGS data of the 20 resequenced *Ae. umbellulata* accessions

Based on the sequence similarity of cloned *Lr9* gene (Wang *et al.*, 2023), we filter out the reads from 20 *Ae. umbellulata* accessions mapped on the *Lr9* region in *Ae. umbellulata* assembly

and visualized using JBrowse2 (Diesh *et al.*, 2023). The filtered reads from the exonic region of *Lr9* were assembled using Trinity (Grabherr *et al.*, 2011) (v2.15.1) and completeness of assembled *Lr9* alleles was checked using web based blastn. The assembled *Lr9* haplotypes were further translated to proteins using a web-based tool Expasy (https://web.expasy.org/translate/). Further, to check for the amino acid substitutions in LR9 protein identified in resequenced accessions compared to cloned LR9, online blastp comparison was conducted and based on the amino acid changes, we divided the LR9 proteins in different haplotypes.

Competing interests

The authors declare no competing interests.

Data availability statement

The genome assembly and raw sequencing data are uploaded to National Center for Biotechnology Information under BioProject

PRJNA1071595. The assembly is also available at https://genomevolution.org/CoGe/GenomeInfo.pl?gid=66166. The seed of *Ae. umbellulata* accessions can be requested by directly contacting the corresponding author UG. The custom scripts used to analyse the data in this manuscript are available at https://github.com/NDSUrustlab/aumb_genome_seq.

References

Abrouk, M., Wang, Y., Cavalet-Giorsa, E., Troukhan, M., Kravchuk, M. and Krattinger, S.G. (2023) Chromosome-scale assembly of the wild wheat relative *Aegilops umbellulata*. *Scientific Data* **10**, 739.

Ankori, H. and Zohary, D. (1962) Natural hybridization between *Aegilops sharonensis* and *Ae. longissima*: a morphological and cytological study. *Cytologia* **27**, 314–324.

Athwal, R.S. and Kimber, G. (1972) The pairing of an alien chromosome with homoeologous chromosomes of wheat. *Can. J. Genet. Cytol.* **14**, 325–333.

Aury, J.M., Engelen, S., Istace, B., Monat, C., Lasserre-Zuber, P., Belser, C., Cruaud, C. *et al.* (2022) Long-read and chromosome-scale assembly of the hexaploid wheat genome achieves high resolution for research and breeding. *Gigascience* **11**, giac034.

Badaeva, E.D. (2002) Evaluation of phylogenetic relationships between five polyploid Aegilops L. species of the U-genome cluster by means of chromosome analysis. *Russian Journal of Genetics* **38**, 664–675.

Badaeva, E.D., Amosova, A.V., Samatadze, T.E., Zoshchuk, S.A., Shostak, N.G., Chikida, N.N., Zelenin, A.V. *et al.* (2004) Genome differentiation in Aegilops. 4. Evolution of the U-genome cluster. *Plant Systematics and Evolution* **246**, 45–76.

Balachandran, P., Walawalkar, I.A., Flores, J.I., Dayton, J.N., Audano, P.A. and Beck, C.R. (2022) Transposable element-mediated rearrangements are prevalent in human genomes. *Nat. Commun.* **13**, 7115.

Bandi, V. and Gutwin, C. (2020) Interactive exploration of genomic conservation. In *Proceedings of Graphics Interface*, pp. 74–83. Toronto: University of Toronto.

Bansal, M., Kaur, S., Dhaliwal, H.S., Bains, N.S., Bariana, H.S., Chhuneja, P. and Bansal, U.K. (2017) Mapping of *Aegilops umbellulata*-derived leaf rust and stripe rust resistance loci in wheat. *Plant Pathol.* **66**, 38–44.

Bansal, M., Adamski, N.M., Toor, P.I., Kaur, S., Molnár, I., Holušová, K., Vrána, J. *et al.* (2020) *Aegilops umbellulata* introgression carrying leaf rust and stripe rust resistance genes Lr76 and Yr70 located to 9.47-Mb region on 5DS telomeric end through a combination of chromosome sorting and sequencing. *Theor. Appl. Genet.* **133**, 903–915.

Bariah, I., Keidar-Friedman, D. and Kashkush, K. (2020) Where the wild things are: transposable elements as drivers of structural and functional variations in the wheat genome. *Front. Plant Sci.* **11**, 585515.

Bickhart, D.M., Rosen, B.D., Koren, S., Sayre, B.L., Hastie, A.R., Chan, S., Lee, J. *et al.* (2017) Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* **49**, 643–650.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O. and Shendure, J. (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125.

Cantarel, B.L., Korf, I., Robb, S.M., Parra, G., Ross, E., Moore, B. *et al.* (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196.

Chen, Y., Zhang, Y., Wang, A.Y., Gao, M. and Chong, Z. (2021) Accurate long-read de novo assembly evaluation with Inspector. *Genome Biol.* **22**, 1–21.

Cheng, H., Jarvis, E.D., Fedrigo, O., Koepfli, K.P., Urban, L., Gemmell, N.J. and Li, H. (2022) Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.* **40**, 1332–1335.

Chhuneja, P., Kaur, S., Goel, R.K., Aghaee-Sarbarzeh, M. and Dhaliwal, H.S. (2007) Introgression of leaf rust and stripe rust resistance genes from *Aegilops umbellulata* to hexaploid wheat through induced homoeologous pairing. In *Wheat Production in Stressed Environments: Proceedings of the 7th International Wheat Conference, Argentina*, pp. 83–90. Dordrecht: Springer Netherlands.

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

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A. *et al.* (2021) Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008.

Daron, J., Glover, N., Pingault, L., Theil, S., Jamilloux, V., Paux, E. *et al.* (2014) Organization and evolution of transposable elements along the bread wheat chromosome 3B. *Genome Biol.* **15**, 1–15.

Diesh, C., Stevens, G.J., Xie, P., De Jesus Martinez, T., Hershberg, E.A., Leung, A. *et al.* (2023) JBrowse 2: a modular genome browser with views of synteny and structural variation. *Genome Biol.* **24**, 74.

Edae, E.A. and Rouse, M.N. (2019) Bulked segregant analysis RNA-seq (BSR-Seq) validated a stem resistance locus in *Aegilops umbellulata*, a wild relative of wheat. *PLoS One* **14**, e0215492.

Edae, E.A., Olivera, P.D., Jin, Y., Poland, J.A. and Rouse, M.N. (2016) Genotype-by-sequencing facilitates genetic mapping of a stem rust resistance locus in *Aegilops umbellulata*, a wild relative of cultivated wheat. *BMC Genomics* **17**, 1–10.

Eldarov, M., Aminov, N. and van Slageren, M. (2015) Distribution and ecological diversity of *Aegilops* L. in the greater and lesser caucasus regions of Azerbaijan. *Genetic Resources and Crop Evolution* **62**, 265–273.

Emms, D.M. and Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14.

Frichot, E. and François, O. (2015) LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution* **6**, 925–929.

Friebe, B., Jiang, J., Tuleen, N. and Gill, B.S. (1995) Standard karyotype of Triticum umbellulatum and the characterization of derived chromosome addition and translocation lines in common wheat. *Theor. Appl. Genet.* **90**, 150–156.

Friskop, A., Green, A., Ransom, J., Liu, Z., Knodel, J., Hansen, B., Halvorson, J. *et al.* (2023) Increase of bacterial leaf streak in hard red spring wheat in North Dakota and yield loss considerations. *Phytopathology* **113**, 2103–2109.

Gill, B.S., Sharma, H.C., Raupp, W.J., Browder, L.E., Hatchett, J.H., Harvey, T.L. *et al.* (1985) Evaluation of Aegilops species for resistance to wheat powdery mildew, wheat leaf rust, Hessian fly, and greenbug. *Plant disease* **69**, 314–316.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652.

Griffiths, S., Sharp, R., Foote, T.N., Bertin, I., Wanous, M., Reader, S., Colas, I. *et al.* (2006) Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat. *Nature* **439**, 749–752.

Guo, Y., Betzen, B., Salcedo, A., He, F., Bowden, R.L., Fellers, J.P., Jordan, K.W. *et al.* (2022) Population genomics of Puccinia graminis f. sp. tritici highlights the role of admixture in the origin of virulent wheat rust races. *Nat. Commun.* **13**, 6287.

International Wheat Genome Sequencing Consortium (IWGSC), Appels, R., Eversole, K., Stein, N., Feuillet, C., Keller, B. *et al.* (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, eaar7191.

Kerber, E.R. (1964) Wheat: reconstitution of the tetraploid component (AABB) of hexaploids. *Science* **143**, 253–255.

Kilian, B., Mammen, K., Millet, E., Sharma, R., Graner, A., Salamini, F. *et al.* (2011) Aegilops. In *Wild Crop Relatives: Genomic and Breeding Resources* (Kole, C., ed), pp. 1–76. Berlin, Heidelberg: Springer Berlin Heidelberg.

Kimber, G. and Yen, Y. (1989) Hybrids involving wheat relatives and autotetraploid Triticum umbellulatum. *Genome* **32**, 1–5.

Letunic, I. and Bork, P. (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**(W1), W293–W296.

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

Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595.

Li, P., Quan, X., Jia, G., Xiao, J., Cloutier, S. and You, F.M. (2016) RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genomics* **17**, 1–10.

Li, L.F., Zhang, Z.B., Wang, Z.H., Li, N., Sha, Y., Wang, X.F., Ding, N. *et al.* (2022) Genome sequences of five Sitopsis species of *Aegilops* and the origin of polyploid wheat B subgenome. *Mol. Plant* **15**, 488–503.

Liu, Z., Zhang, X., Wan, Y., Liu, K. and Wang, D. (2002) Characterization of high-molecular-weight glutenin subunits and their coding genes from *Aegilops umbellulata*. *J. Integr. Plant Biol.* **44**, 809.

Manni, M., Berkeley, M.R., Seppey, M. and Zdobnov, E.M. (2021) BUSCO: assessing genomic data quality and beyond. *Current Protocols* **1**, e323.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K. *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303.

Molnár, I., Vrána, J., Burešová, V., Cápal, P., Farkas, A., Darkó, É., Cseh, A. *et al.* (2016) Dissecting the U, M, S and C genomes of wild relatives of bread wheat (*Aegilops* spp.) into chromosomes and exploring their synteny with wheat. *Plant J.* **88**, 452–467.

Okada, M., Yoshida, K., Nishijima, R., Michikawa, A., Motoi, Y., Sato, K. and Takumi, S. (2018) RNA-seq analysis reveals considerable genetic diversity and provides genetic markers saturating all chromosomes in the diploid wild wheat relative *Aegilops umbellulata*. *BMC Plant Biol.* **18**, 1–13.

Olivera, P., Newcomb, M., Szabo, L.J., Rouse, M., Johnson, J., Gale, S., Luster, D.G. *et al.* (2015) Phenotypic and genotypic characterization of race TKTTF of *Puccinia graminis* f. sp. *tritici* that caused a wheat stem rust epidemic in Southern Ethiopia in 2013-14. *Phytopathology* **105**, 917–928.

Olivera, P.D., Sikharulidze, Z., Dumbadze, R., Szabo, L.J., Newcomb, M., Natsarishvili, K., Rouse, M.N. *et al.* (2019) Presence of a sexual population of *Puccinia graminis* f. sp. *tritici* in Georgia provides a hotspot for genotypic and phenotypic diversity. *Phytopathology* **109**, 2152–2160.

Papon, N., Lasserre-Zuber, P., Rimbert, H., De Oliveira, R., Paux, E. and Choulet, F. (2023) All families of transposable elements were active in the recent wheat genome evolution and polyploidy had no impact on their activity. *Plant Genome* **16**, e20347.

Peichel, C.L., Sullivan, S.T., Liachko, I. and White, M.A. (2016) Improvement of the threespine stickleback (*Gasterosteus aculeatus*) genome using a Hi-C-based proximity-guided assembly method. *J Hered* **108**(6), 693–700.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics* **81**, 559–575.

Rabanus-Wallace, M.T., Hackauf, B., Mascher, M., Lux, T., Wicker, T., Gundlach, H., Baez, M. *et al.* (2021) Chromosome-scale genome assembly provides insights into rye biology, evolution and agronomic potential. *Nat. Genet.* **53**, 564–573.

Riley, R. and Chapman, V. (1958) Genetic control of the cytologically diploid behaviour of hexaploid wheat. *Nature* **182**, 713–715.

Ruban, A.S. and Badaeva, E.D. (2018) Evolution of the S-genomes in *Triticum-Aegilops* alliance: evidences from chromosome analysis. *Front. Plant Sci.* **9**, 1756.

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S. *et al.* (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274**, 765–768.

Sapkota, S., Zhang, Q., Chittem, K., Mergoum, M., Xu, S.S. and Liu, Z. (2018) Evaluation of triticale accessions for resistance to wheat bacterial leaf streak caused by *Xanthomonas translucens* pv. undulosa. *Plant Pathol.* **67**, 595–602.

Sasanuma, T., Chabane, K., Endo, T.R. and Valkoun, J. (2004) Characterization of genetic variation in and phylogenetic relationships among diploid *Aegilops* species by AFLP: incongruity of chloroplast and nuclear data. *Theor. Appl. Genet.* **108**, 612–618.

Schneider, A., Molnár, I. and Molnár-Láng, M. (2008) Utilisation of *Aegilops* (goatgrass) species to widen the genetic diversity of cultivated wheat. *Euphytica* **163**, 1–19.

Sears, E.R. (1956) The transfer of leaf-rust resistance from *Aegilops umbellulata* to wheat. *Brookhaven Symposium in Biol.* **9**, 1–22.

Smit, A.F.A., Hubley, R. and Green, P. (1996) RepeatMasker Open-3.0. https://www.repeatmasker.org/

Song, Z., Dai, S., Bao, T., Zuo, Y., Xiang, Q., Li, J., Liu, G. *et al.* (2020) Analysis of structural genomic diversity in *Aegilops umbellulata*, Ae. markgrafii, Ae. comosa, and Ae. uniaristata by fluorescence in situ hybridization karyotyping. *Front. Plant Sci.* **11**, 710.

Song, Z.P., Zuo, Y.Y., Xiang, Q., Li, W.J., Jian, L.I., Liu, G. *et al.* (2023) Investigation of *Aegilops umbellulata* for stripe rust resistance, heading date, and the contents of iron, zinc, and gluten protein. *J. Integr. Agric.* **22**, 1258–1265.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313.

Stanke, M. and Morgenstern, B. (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**(suppl_2), W465–W467.

Szabo, L.J., Olivera, P.D., Wanyera, R., Visser, B. and Jin, Y. (2022) Development of a diagnostic assay for differentiation between genetic groups in clades I, II, III, and IV of *Puccinia graminis* f. sp. *tritici*. *Plant Dis.* **106**, 2211–2220.

UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**(D1), D480–D489.

Vaillancourt, B. and Buell, C.R. (2019) *High molecular weight DNA isolation method from diverse plant species for use with Oxford Nanopore Sequencing*. BioRxiv, 783159.

Wang, Y., Tang, H., DeBarry, J.D., Tan, X., Li, J., Wang, X., Lee, T.H. *et al.* (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49.

Wang, J., Wang, C., Zhen, S., Li, X. and Yan, Y. (2018) Low-molecular-weight glutenin subunits from the 1U genome of *Aegilops umbellulata* confer superior dough rheological properties and improve breadmaking quality of bread wheat. *J. Sci. Food Agric.* **98**, 2156–2167.

Wang, W., Ji, W., Feng, L., Ning, S., Yuan, Z., Hao, M., Zhang, L. *et al.* (2022) Characterization of novel low-molecular-weight glutenin subunit genes from the diploid wild wheat relative *Aegilops umbellulata*. *Plant Genetic Resources* **20**, 1–6.

Wang, Y., Abrouk, M., Gourdoupis, S., Koo, D.H., Karafiátová, M., Molnár, I., Holušová, K. *et al.* (2023) An unusual tandem kinase fusion protein confers leaf rust resistance in wheat. *Nat. Genet.* **55**, 914–920.

Wicker, T., Gundlach, H., Spannagl, M., Uauy, C., Borrill, P., Ramírez-González, R.H. *et al.* (2018) Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biol.* **19**, 1–18.

Wicker, T., Stritt, C., Sotiropoulos, A.G., Poretti, M., Pozniak, C., Walkowiak, S., Gundlach, H. *et al.* (2022) Transposable element populations shed light on the evolutionary history of wheat and the complex co-evolution of autonomous and non-autonomous retrotransposons. *Adv. Genet.* **3**, 2100022.

Yang, Y.C., Tuleen, N.A. and Hart, G.E. (1996) Isolation and identification of *Triticum aestivum* L. em. Thell. cv Chinese Spring-*T. peregrinum* Hackel disomic chromosome addition lines. *Theor. Appl. Genet.* **92**, 591–598.

Zhang, H., Jia, J., Gale, M.D. and Devos, K.M. (1998) Relationships between the chromosomes of *Aegilops umbellulata* and wheat. *Theor. Appl. Genet.* **96**, 69–75.

Zhu, Z., Zhou, R., Kong, X., Dong, Y. and Jia, J. (2006) Microsatellite marker identification of a *Triticum aestivum-Aegilops umbellulata* substitution line with powdery mildew resistance. *Euphytica* **150**, 149–153.

Zhukovsky, P.M. (1928) Kritiko-systematischeskii obzor vydov roda *Aegilops* L. (Specierum generis *Aegilopis* L. revisio critica). *Trudy Prikl Bot* **18**(1), 417–609 (in Russian with English summary on pp 584–609).

Zimin, A.V., Puiu, D., Luo, M.C., Zhu, T., Koren, S., Marçais, G., Yorke, J.A. *et al.* (2017) Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res.* **27**, 787–792.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Methods S1 Disease phenotyping.

Figure S1 (a) Hi-C contact map showing the intrachromosomal interaction heatmap in the assembled chromosomes of *Aegilops umbellulata*; (b) graphical presentation of *Aegilops umbellulata* chromosomes (blue bars) with telomeres (red) and chromosome length; (c) BUSCO assessments for analysing the quality of assembled genome.

Figure S2 Types of transposable elements (TEs) identified in *Aegilops umbellulata* in comparison to three sub-genomes of wheat (ref v2.1). (a) Proportion of class-I, class-II, and unclassified

TEs; (b) proportion of sub-classes within class-I TEs; and (c) proportion of sub-classes of class-II TEs.

Figure S3 Comparative genome analysis of *Aegilops umbellulata* with *Brachypodium distachyon* (a), *Hordeum vulgare* (b), A-B-D sub-genomes of *Triticum aestivum* (c), *Triticum urartu* (A genome) (d), and *Aegilops speltoides* (B genome) (e).

Figure S4 Syntenic relationship of *Aegilops umbellulata* (au) with *Triticum urartu* (tu), *Aegilops speltoides* (as), and *Aegilops sharonensis* (sh).

Figure S5 Syntenic relationship of individual *Aegilops umbellulata* (au) chromosomes with *Aegilops tauschii* (at) and *Aegilops longissima* (al).

Figure S6 Disease reaction of *Aegilops umbellulata* accessions and susceptible wheat cultivars (Prosper, Thatcher, and Morocco) to *Puccinia triticina* races, (a) TNBJS and (b) MNPSD.

Figure S7 Bacterial leaf streak (BLS) resistance in *Aegilops umbellulata* accession, PI 554417 compared to sequenced accession, PI 554389 against BLS-P3 isolate at seedling stage.

Figure S8 The image illustrates diverse spike structures within *Aegilops umbellulata*, showcasing the morphological variations in spike architecture, a crucial trait for understanding the genetic diversity and evolutionary dynamics in this plant species.

Figure S9 Morphological variations for plant architecture and flowering in *Aegilops umbellulata*, provides insights into the morphological variability within this species.

Figure S10 Pairwise nucleotide diversity (?) analysis of resequenced *Aegilops umbellulata* accessions compared to sequenced PI 554389 for each chromosome.

Figure S11 Multiple sequence alignment of cloned *Lr9* gene from TA1851 and six Lr9 haplotypes (Hap) found in 20 *Aegilops umbellulata* accessions.

Table S1 Raw data yield of PacBio high-fidelity (HiFi) circular consensus sequencing (CCS).

Table S2 Raw data yield of Nanopore MinION sequencing.

Table S3 Summary statistics of genome assembly using PacBio HiFi reads.

Table S4 Summary statistics of scaffolds after high-throughput chromosome conformation capture (Hi-C).

Table S5 Summary of manual gap closing of seven pseudomolecules using PacBio HiFi and ONT ultra-long reads.

Table S6 Synteny guided chromosome assignment to pseudomolecules based on IWGSC RefSeq_v2.1 D sub-genome.

Table S7 Summary of chromosomes length, genomic content, and predicted gene models of *Aegilops umbellulata* genome assembly.

Table S8 Comparison of GC content (%) of *Aegilops umbellulata* and wheat reference genome.

Table S9 Genome completeness of *Aegilops umbellulata* based on BUSCO score.

Table S10 Type and number of predicted resistance gene analogs (RGAs) in *Aegilops umbellulata* and *Triticum aestivum* reference (v2.1) genome.

Table S11 Proportion of Transposable Elements (TEs) at whole genome level for *Aegilops umbellulata* and *Triticum aestivum* reference (v2.1) genome.

Table S12 High and Low confidence gene models for different species used in phylogenetic analysis.