

A pangenome of tetraploid wheat reveals the genetic architecture underlying domestication and genomic diversity for breeding

Received: 5 December 2025

Accepted: 16 June 2026

Published online: 22 July 2026

 Check for updates

A list of authors and their affiliations appears at the end of the paper

Tetraploid wheat (*Triticum turgidum* L., BBAA), a key pasta crop, serves as an untapped genetic resource with rich genomic diversity for hexaploid bread wheat improvement. Here we de novo assembled 12 genomes spanning all 10 recognized tetraploid wheat (genome BBAA) subspecies, and a graph-based pangenome was constructed. Chromosome rearrangements drove subgenome asymmetry and shaped genomic divergence, with an average of 0.25 million structural variations per accession, predominantly attributable to transposon activity. Using 736 globally distributed tetraploid wheat accessions, we identified locally adapted subgroups with untapped breeding potential and discovered a novel retrotransposon-induced loss-of-function *Btr1-A* allele responsible for convergent adaptation of non-brittle rachis. Genome-wide association studies identified 287 loci associated with 32 traits. A homeodomain-leucine zipper transcription factor HAT14-B that enhances both spikelet number and grain size was identified. This subspecies-wide pangenome enriches Triticeae AB subgenome resources and facilitates the discovery and application of agronomically important genetic variations.

Tetraploid wheat (*Triticum turgidum* L., BBAA, $2n = 4x = 28$) contributes two subgenomes (B and A) to the origin of hexaploid bread wheat (*Triticum aestivum* L., BBAADD, $2n = 6x = 42$, also known as common wheat), and serves as a valuable genetic reservoir for its improvement^{1,2}. Domesticated tetraploid wheat is a primary crop for non-bread-wheat-based food products, including pasta, bulgur and couscous, with an annual production of 37–40 million tons (refs. 3,4). Tetraploid wheat (*T. turgidum* L.) encompasses a spectrum of breeding-improved cultivars, domesticated landraces, as well as wild types⁵, which are absent from bread wheat. This diversity makes tetraploid wheat an ideal system for investigating the genetic basis of wheat domestication.

Tetraploid wheat (*T. turgidum* L.) is thought to have originated from the Near East⁶ about 800,000 years ago⁷ and consists of ten distinct subspecies distributed across diverse geographical regions. The evolutionary history of tetraploid wheat is approximately 80 times longer than that

of bread wheat⁷, and has much greater genetic diversity⁸. Numerous key functional genes associated with breeding values have been identified in tetraploid wheat, and exploited for breeding and genetically validated in bread wheat^{9–11}. Exploring untapped genomic resources in tetraploid wheat can broaden the genetic basis of bread wheat and advance future breeding amid rising global food demand. Though available genomes of wild emmer wheat (*Triticum dicoccoides*) Zavitan¹², durum wheat cultivar Svevo¹³ and Langdon genome¹⁴ have facilitated key findings, a comprehensive graph-based pangenome and a high-resolution variation map remain essential for efficient identification of allelic variants and candidate genes for bread wheat improvement.

Here we de novo assembled 12 tetraploid wheat genomes, covering all 10 known subspecies², to establish a comprehensive graph-based pangenome. This genome panel facilitated the investigation of genomic variations and evolutionary relationships, including genes, transposable

✉ e-mail: shisheng.chen@pku-iaas.edu.cn; qing.sang@pku-iaas.edu.cn; deng@pku.edu.cn; mao_dehu@nwsuaf.edu.cn; small@nwsuaf.edu.cn; baoxing.song@pku-iaas.edu.cn

Table 1 | Assembly and annotation statistics for the 12 tetraploid wheat genomes

Accession name	Type	Subspecies	Assembly size (Gb)	Contig N50 (Mb)	Repeat (%)	Genome BUSCO (%)	QV	LAI	No. of HC genes
Kronos	Free-threshing tetraploid wheat	<i>Triticum durum</i>	10.82	172.76	91.67	99.0	66.02	21.79	85,172
XM001097	Free-threshing tetraploid wheat	<i>Triticum turanicum</i>	10.49	38.24	91.68	98.9	71.19	18.72	85,640
NU00021	Free-threshing tetraploid wheat	<i>Triticum carthlicum</i>	10.50	48.26	91.42	99.0	71.07	18.67	84,714
IG77365	Free-threshing tetraploid wheat	<i>Triticum aethiopicum</i>	10.45	37.87	91.50	99.1	69.97	18.65	84,531
XM000719	Free-threshing tetraploid wheat	<i>Triticum polonicum</i>	10.65	46.46	91.28	99.0	69.61	18.81	89,423
XM000834	Free-threshing tetraploid wheat	<i>Triticum turgidum</i>	10.51	49.32	91.56	98.9	67.34	18.88	84,608
PI294478	Domesticated tetraploid wheat	<i>Triticum ispahanicum</i>	10.51	43.62	91.56	99.0	70.51	18.68	85,045
PI330553	Domesticated tetraploid wheat	<i>Triticum karamyshevii</i>	10.47	38.44	92.97	99.0	68.15	19.53	78,276
IG99236	Domesticated tetraploid wheat	<i>Triticum dicoccon</i>	10.46	35.79	91.41	99.0	70.42	19.11	84,451
NU01905	Wild emmer wheat	<i>Triticum dicoccoides</i>	10.61	46.26	91.44	99.0	70.19	18.67	84,777
NU01954	Wild emmer wheat	<i>Triticum dicoccoides</i>	10.56	49.81	91.48	99.1	69.22	18.81	83,408
Zavitan	Wild emmer wheat	<i>Triticum dicoccoides</i>	10.51	41.17	92.95	99.0	69.15	18.81	78,379

elements (TEs) and structural variants (SVs). Using 736 tetraploid wheat accessions, we constructed a genetic-variation atlas and detected 287 loci underlying 32 agronomic traits via genome-wide association studies (GWAS). The graph-based pangenome offers a fundamental platform for identifying elite genes such as *HAT14* for wheat improvement.

This extensive genomic resource not only provides new insights into the evolutionary and genetic architecture of tetraploid wheat but also offers a powerful foundation for the genetic improvement of wheat and other Triticeae crops, with potentially broad implications for global food security and sustainable agriculture.

Results

High-quality genome assemblies and annotations

To construct a tetraploid wheat pangenome, we collected 12 accessions covering all 10 known tetraploid wheat (*T. turgidum* L.) subspecies (Table 1 and Supplementary Table 1). Phenotypes of the 12 accessions are diverse, including plant height, tiller angle, seed size and shape, awn length and spike morphology (Fig. 1a–c). For durum cultivar (cv.) Kronos, we assembled a highly contiguous genome with a contig N50 of 172.76 megabases (Mb) using 998.37 gigabases (Gb) of PacBio HiFi reads, 576.75 Gb of Oxford Nanopore Technology (ONT) reads (length ≥ 100 kilobases (kb)) and 1.04 terabases (Tb) of Hi-C reads (Supplementary Fig. 1 and Supplementary Note 1). On the basis of contig N50 saturation analysis (Supplementary Fig. 2), we produced an average of 556.46-Gb PacBio HiFi and 439.39-Gb Hi-C reads for de novo assembly of the other 11 tetraploid wheat accessions. Contig N50 scores for these 11 genomes were 35.79–49.81 Mb, averaging 43.2 Mb (Fig. 1d) and genome length ranged between 10.45 Gb and 10.65 Gb (Supplementary Table 1). All 12 accessions were anchored to the chromosome level using Hi-C reads, achieving an average anchoring rate of 98.81% (Table 1 and Supplementary Fig. 3). Meanwhile, an average of 23 telomeres for each assembly were identified, of which accession IG99236 (*Triticum dicoccon*) has the most detectable telomeres ($n = 27$), whereas accession XM000834 (*T. turgidum*) has the least ($n = 20$; Fig. 1e). These assemblies achieved a benchmarking universal single-copy orthologs (BUSCO) completeness score of 99%, a quality value (QV) of 69.41 and a long terminal repeat assembly index (LAI) of

19.19, demonstrating high quality and contiguity. Coupled with good synteny with published assemblies^{12–14}, these metrics indicate the high quality and high contiguity of the genomic resources (Table 1, Supplementary Figs. 4–6 and Supplementary Note 2).

TEs account for 90.92–92.81% of these tetraploid wheat genomes, with long terminal repeat retrotransposons (LTR-RTs) as the predominant class (Fig. 1f and Supplementary Table 2). Centromere regions across all 12 accessions were identified by CENH3 chromatin immunoprecipitation and sequencing (ChIP-seq)¹⁵ and confirmed by repeat-sequence analysis (Supplementary Table 3 and Supplementary Figs. 7 and 8), ranging from 6.38 Mb (chromosome 5B of IG99236) to 20.47 Mb (chromosome 6A of Zavitan) in length. Using ab initio prediction, homology-search and transcriptome-based evidence, 78,276 to 89,423 high-confidence (HC) predicted genes (with an average functional annotation rate of 94.17%) were identified for the 12 new assemblies as well as the published Svevo genome¹³ (Table 1 and Supplementary Tables 4 and 5). Collectively, these genome assemblies represent a high-quality and comprehensive genomic resource for tetraploid wheat.

Pangenome construction and analysis

Presence/absence variation (PAV) of genes can result in variable phenotypes. We evaluated accession representativeness by profiling core and pangenome gene dynamics with incremental genome addition, confirming that the tetraploid wheat pangenome approached a plateau (Fig. 2a). We clustered the genes and defined 50,787 distinct gene families. Among them, 26,267 core gene families (51.7%, 64,893 genes per accession) existed in all 13 accessions; 3,770 softcore gene families (7.4%, 6,077 genes per accession) were present in 12 accessions; 19,879 dispensable gene families (39.1%, 11,383 genes per accession) occurred in 2–11 accessions; and 871 private gene families (1.7%, 147 genes per accession) were specific to a single accession (Fig. 2b,c).

Furthermore, 90.81% of the core genes had conserved protein domains, much higher than that of the softcore (88.67%), dispensable (85.11%) and private genes (85.27%) (Supplementary Fig. 9a). A K_a/K_s analysis also showed that core (median value of K_a/K_s 0.28) genes underwent stronger purifying selection than softcore (median

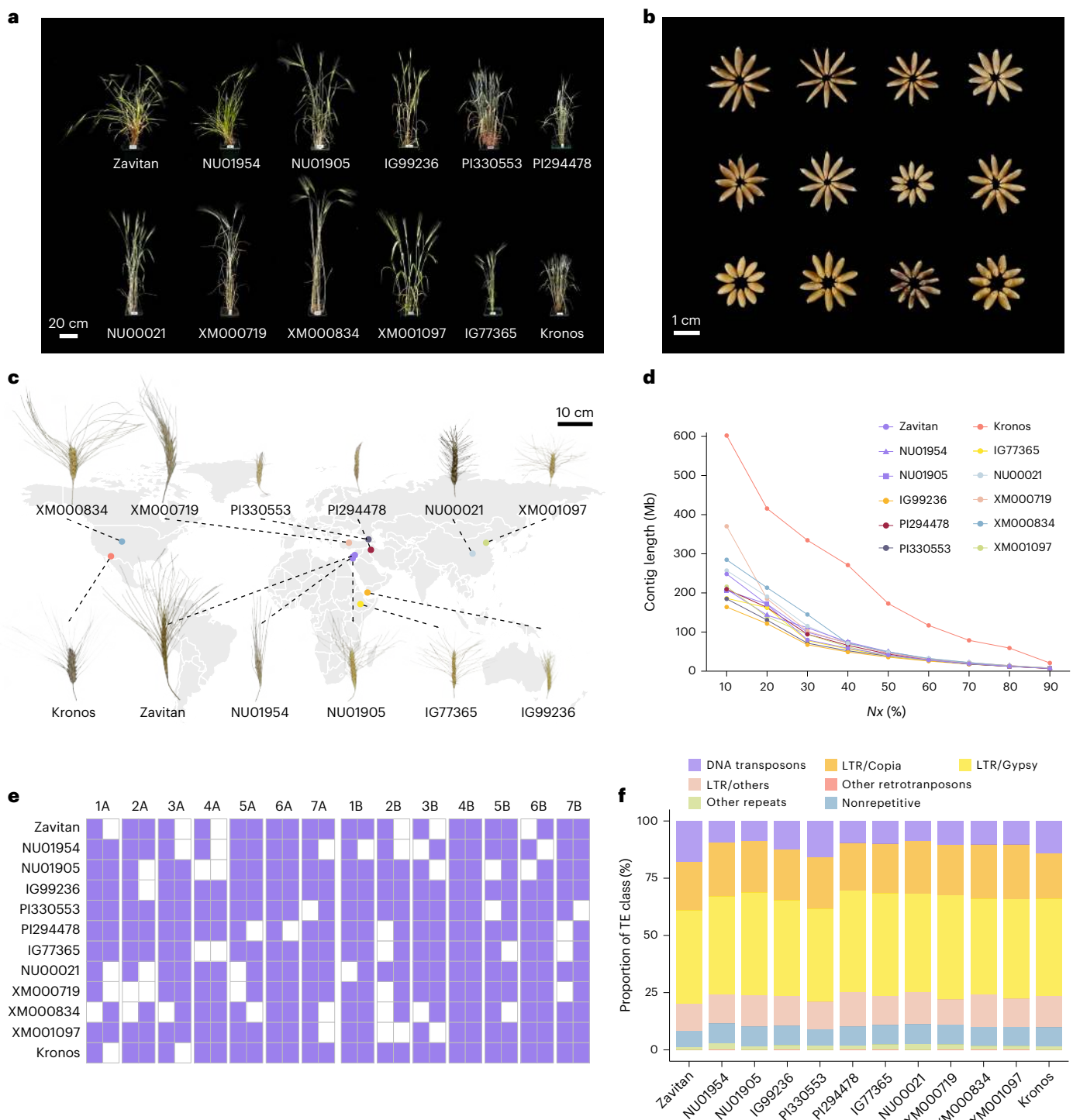


Fig. 1 | Phenotypes and genome characterization. **a**, Plant structures of the 12 sequenced tetraploid wheat accessions. **b**, Seed morphology of the 12 sequenced accessions. **c**, Geographical distribution and spike morphology. **d**, Evaluation of the continuities for the 12 genome assemblies. N_x (%), length of the shortest contig in the group of the longest contigs that represent $x\%$ of the assembly length. **e**, Presence and absence of telomeres in each genome assembly.

Purple squares indicate the presence of telomeres and white squares represent their absence. **f**, Proportion of annotated TEs in each genome. LTR/unknown was included in LTR/others class. The map in **c** was plotted using the R package *rworldmap* (GPL v.2 license, <https://github.com/AndySouth/rworldmap>), with base map data from Natural Earth (<http://www.naturalearthdata.com>).

value of K_a/K_s 0.36), dispensable (median value of K_a/K_s 0.42) and private (median value of K_a/K_s 0.51) genes (Supplementary Fig. 9b). In addition, TE densities within and flanking private, dispensable and softcore genes were higher than that for core genes, indicating that the PAV of these genes were more likely to be related to TE activity^{16,17} (Fig. 2d and Supplementary Fig. 9c). Enrichment analysis

showed that core genes were over-represented for certain fundamental biological functions, such as ATP binding (GO:0005524), DNA integration (GO:0015074) and protein glycosylation (GO:0006486). In contrast, softcore, dispensable and private genes were enriched in specific pathways, including root development (GO:0048364), shoot system development (GO:0048367), response to oxidative stress

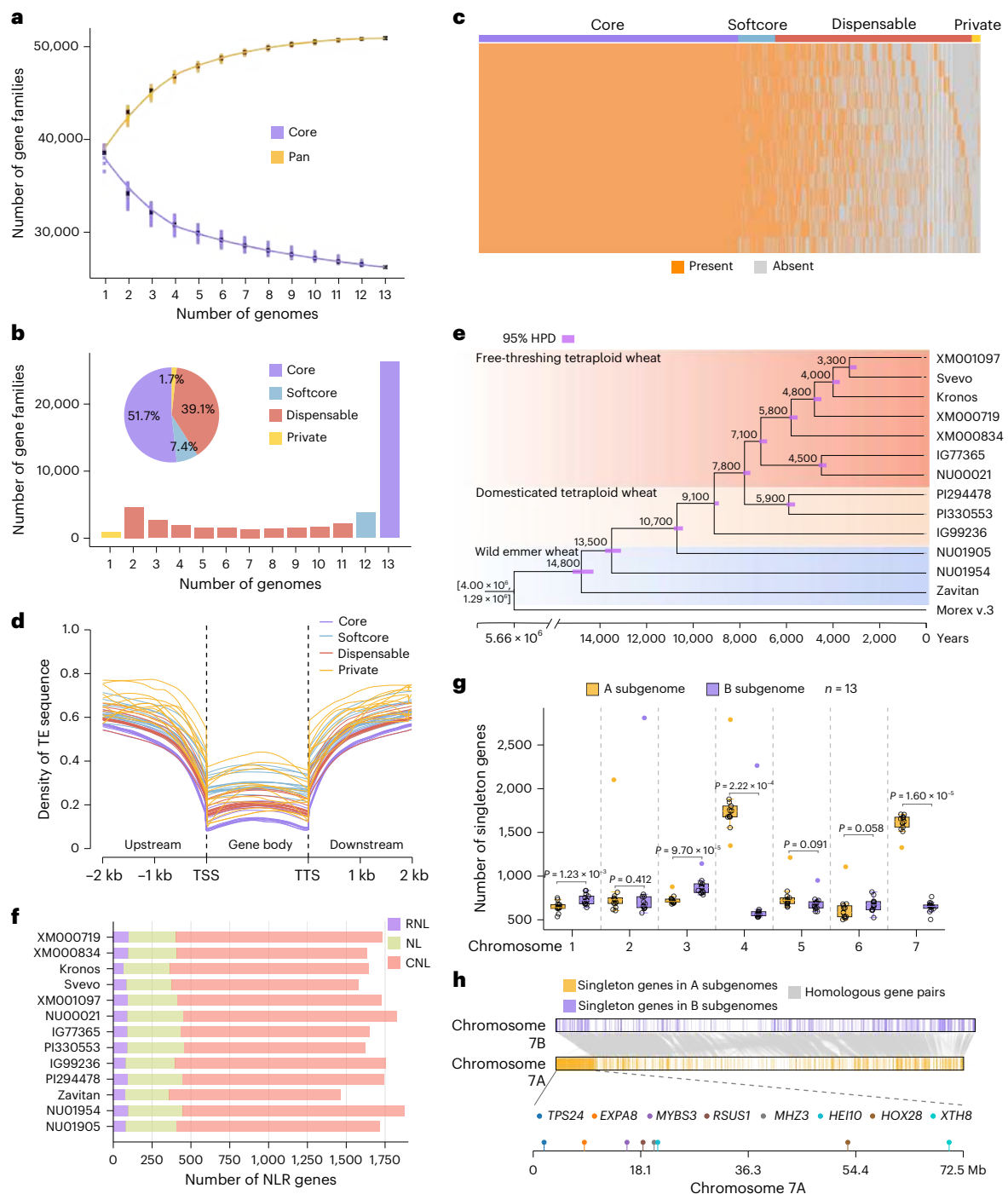


Fig. 2 | Pangenome analysis of 13 tetraploid wheat accessions. a, Pan-core gene family curve in representatives. **b**, Numbers of core, softcore, dispensable and private gene families identified in the pangenome. The histogram shows the frequency of gene families and the pie chart depicts the proportions of core, softcore, dispensable and private gene families. **c**, Presence and absence information of all gene families. **d**, Density of TE sequences in gene bodies and 2-kb flanking regions. **e**, Phylogenetic tree of 13 tetraploid wheat accessions with barley (Morex v.3) as the outgroup. The timeline at the bottom is labeled in descending order, marking key time points from 5.66 million years ago (Ma) to present. **f**, Gene numbers of three NLR subfamilies in 13 tetraploid wheats.

g, Number of singleton genes in each genome. Statistical tests were conducted using two-sided Mann–Whitney *U*-test, with a Bonferroni-corrected significance threshold ($n = 13$). Center line, median; box lower and upper edges, 25% and 75% quartiles, respectively; whiskers, 1.5× interquartile range (IQR); colored dots, outliers. **h**, Singleton genes on chromosomes 7A and 7B in cv. Kronos. CNL, coiled-coil-domain-containing CC–NBS–LRR-type genes; HPD, highest posterior density interval; NL, NL-type gene that contains at least one NBS (nucleotide-binding site) and LRR (leucine-rich repeat) domain; RNL, RESISTANCE TO POWDERY MILDEW 8 (RPW8)-domain-containing RPW8–NBS–LRR-type genes.

(GO:0034599), recognition of pollen (GO:0048544), endosperm development (GO:0009960) and cadmium ion transmembrane transporter activity (GO:0015086) (Supplementary Fig. 10). This analysis reveals extensive diversity in the gene space of tetraploid wheat,

providing a foundation for dissecting the genetic architecture underlying trait variation.

Pangenomes enable tracing of evolutionary histories and genetic divergence among subspecies. Phylogenetic analysis was

first performed using 842 single-copy genes present in 13 tetraploid wheat genomes, with barley (Morex v.3) as the outgroup. All accessions were classified into three major subpopulations, including wild emmer wheat, domesticated tetraploid wheat and free-threshing tetraploid wheat (Fig. 2e). Divergence time estimation revealed that domesticated tetraploid wheat diverged from wild emmer wheat approximately 10,700 years ago and free-threshing tetraploid wheat diverged from domesticated tetraploid wheat around 7,800 years ago (Fig. 2e), consistent with published studies^{18–23}. High-quality de novo assemblies of tetraploid wheat genomes enabled a more detailed investigation of the close ancestor for bread wheat. Integration of 34 representative bread wheat accessions found that the A and B subgenomes of these hexaploids were genetically closer to that of free-threshing tetraploid wheat, with the closest relationships to Persian wheat (*Triticum carthlicum*) and Ethiopian wheat (*Triticum aethiopicum*; Supplementary Fig. 11).

Dynamic expansion and contraction of gene families are crucial evolutionary drivers underlying crop environmental adaptation. We characterized these variations across 13 tetraploid wheat genomes. In the transition from wild emmer to domesticated tetraploid wheat, 561 gene families expanded and 340 contracted. Additionally, 675 gene families exhibited distinct changes between domesticated and free-threshing tetraploid wheat (Supplementary Table 6). Disease-resistance genes generally show substantial changes in copy number^{24,25} and several of these have been identified in tetraploid wheat and genetically validated in bread wheat^{9,11}. Enrichment analysis revealed that disease resistance-related genes were enriched in contracted gene families from domesticated tetraploid wheat to free-threshing tetraploid wheat (Supplementary Fig. 12). We further analyzed nucleotide-binding sites and leucine-rich-repeat^{26,27} (NLR, also known as NBS–LRR) genes. An average of 1,687 NLR genes were identified per genome, wherein coiled-coil domain-containing CC–NBS–LRR-type (CNL) genes were the most abundant, followed by NL genes (only including at least one NBS and LRR domain) and RESISTANCE TO POWDERY MILDEW 8 (RPW8)-domain-containing RPW8–NBS–LRR-type (RNL) genes (Fig. 2f, Supplementary Fig. 13a and Supplementary Table 7). This pattern is similar to that observed in *Panicum hallii* and rice (*Oryza sativa*) but different from the observations in Brassicaceae family species^{26–29}. Additionally, the number of predicted NLR genes in the B subgenomes (56%) was greater than that in A subgenomes (44%).

To investigate the diversity of NLR genes across tetraploid wheat subspecies, we constructed a pan-NLRome containing 308 core, 66 softcore, 435 dispensable and 12 private gene families (Supplementary Figs. 13b,c). The large proportion of private and dispensable NLR genes indicated high diversity among the subspecies. Upon comparing bread wheat NLR genes to the tetraploid pan-NLRomes, 970 gene families were absent from the hexaploid wheat genomes (Supplementary Fig. 13d,e), suggesting that tetraploid wheats harbor highly dynamic NLR gene sets that represent valuable disease-resistance gene resources for hexaploid wheat^{30,31}.

Subgenome asymmetry in tetraploid wheat

Pangenome data enable characterization of asymmetric evolution between subgenomes. We observed marked differences between A and

B tetraploid wheat subgenomes, including chromosome length, TE distribution and gene abundance (Supplementary Fig. 14). To further characterize the asymmetric evolution across all 13 genomes, we identified the reciprocal best hit, single-side best hit and singleton (no corresponding syntenic homologs gene between subgenomes) genes between A and B subgenomes (Supplementary Table 8). On average, each genome contained 12,125 singleton genes (6,962 in A and 5,163 in B) (Fig. 2g). Analysis of singleton distributions revealed differences in gene counts between A and B subgenomes on chromosomes 1, 3, 4 and 7 in both domesticated tetraploid wheat and free-threshing tetraploid wheat, with disparities particularly pronounced on chromosomes 4 and 7. Subgenome-specific singleton genes were enriched in the terminal regions of the long arm of chromosome 4A (4AL) and the short arm of chromosome 7 (7AS), as observed in all the assemblies, suggesting that the 7BS–4AL translocation predates the subspeciation of tetraploid wheat^{32,33} (Supplementary Fig. 15). This ancestral rearrangement likely amplified subgenome divergence by increasing the number of singleton genes. Functional annotation revealed that those singletons include gene for disease resistance, such as *TPS24*, *chi11*, *Pht1* and *SIK1*; and grain development, for example *RSUS1*, *MHZ3*, *HEI10* and *DDM1a* (Fig. 2h and Supplementary Table 9), suggesting that the translocation potentiated subgenome-specific evolution and diversification in polyploid wheat.

Structural variation among tetraploid wheat

We further characterized genomic diversity at base-pair resolution by comparing 12 newly assembled genomes and found an average of 83.51% matched sequences and as high as 12.58% of gap sequences (Supplementary Fig. 16). By visualizing the genome alignments (Fig. 3a), we observed three large karyotype variations, including a translocation between chromosomes 2A and 4B of accession IG77365 (*T. aethiopicum*), a 317.34-Mb inversion on chromosome 5A of accession IG77365³⁴ and a 355.54-Mb inversion on chromosome 7A of accession NU00021 (*T. carthlicum*). These large karyotype variations were verified by fluorescence in situ hybridization and Hi-C heatmaps (Fig. 3b,c, Supplementary Note 3 and Supplementary Figs. 17 and 18). On average, each accession had 233,976 SVs, including 116,990 deletions and 116,986 insertions (Fig. 3d and Supplementary Table 10). The number of SVs located in B subgenomes was larger than in A subgenomes, underscoring asymmetry across subgenomes (Fig. 3e). PCR validation of 30 randomly selected SVs validated the accuracy of these assignments (Supplementary Fig. 19 and Supplementary Table 11).

These SVs account for 11.29% on average of the genome sequences (ranging from 0.78 Gb to 1.77 Gb), with an average length of 3,450 base pairs (bp) (Fig. 3f). We also observed that on average 6.01% of all SVs were located upstream or downstream of protein-coding genes and 3.31% of SVs overlapped with coding regions. These overlapping SVs might have substantial functional relevance by affecting gene structure or expression (Fig. 3g). Identity-by-state (IBS) values based on SVs also showed higher potential lineage homology among free-threshing tetraploid wheat but lower in wild emmer wheat (Fig. 3h). We observed that an average of 94.7% of SVs overlapped with TEs (Fig. 3i), which is consistent with the hypothesis that most SVs are derived from TE activation³⁵. Changes in SV number were also observed among core, softcore

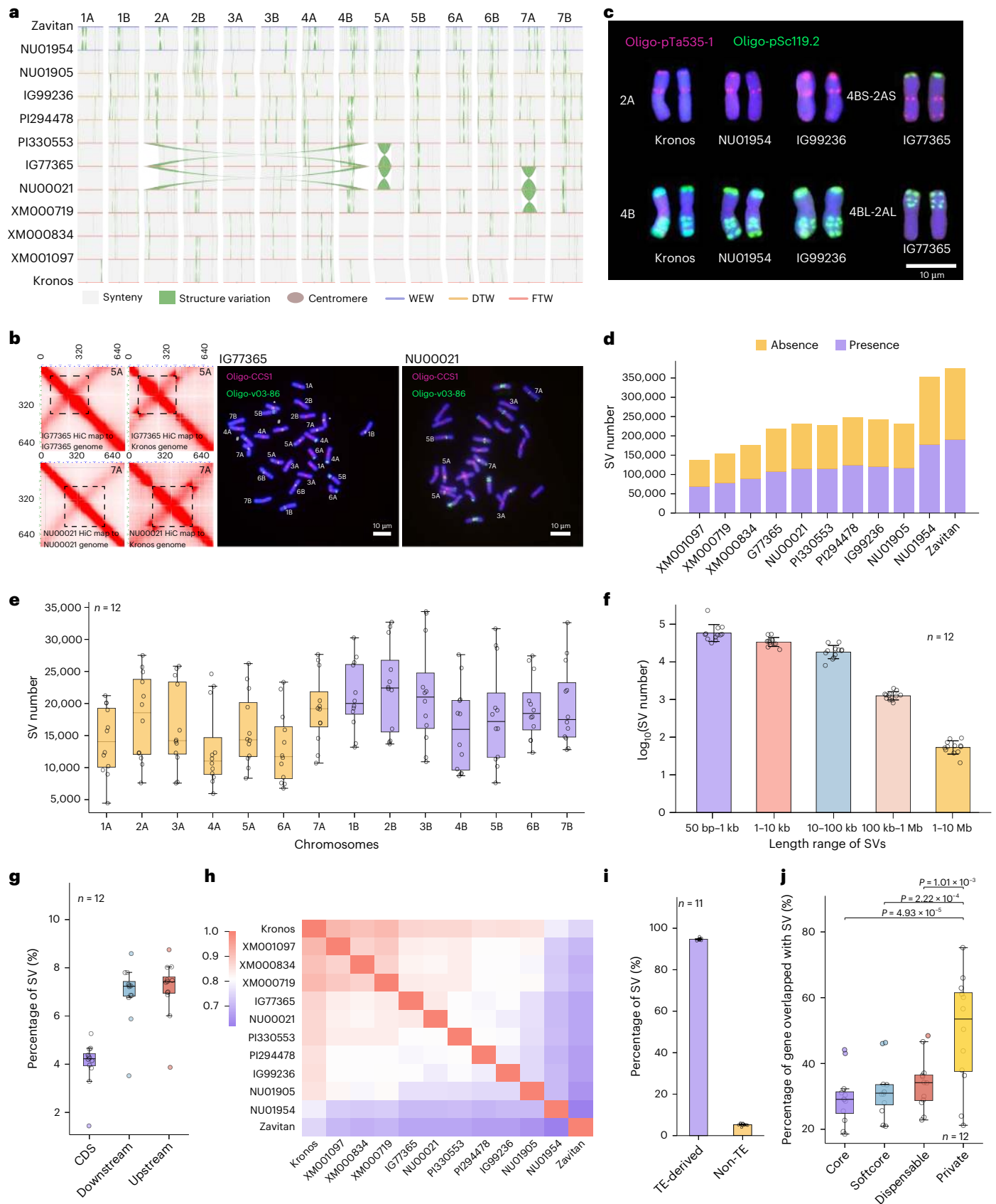
Fig. 3 | Structural variants across tetraploid wheat genomes. **a**, Whole-genome synteny of the 12 newly assembled genomes. **b**, Hi-C (left) and fluorescence in situ hybridization (right) assays confirm the inversion of the near-centromeric region on chromosome 5A of accession IG77365 and on chromosome 7A of accession NU00021. The axes on the heatmap represent megabases. Fluorescence in situ hybridization assays were repeated twice. **c**, Fluorescence in situ hybridization for chromosomes 2A and 4B of accession IG77365. **d**, Number of SVs (deletions and insertions) observed in each accession, using the Kronos genome as a reference. **e**, Number of SVs in each chromosome across 11 newly assembled and Svevo genomes ($n = 12$), with Kronos as reference. **f**, Distribution of SV length across 11 newly assembled and Svevo genomes ($n = 12$), with Kronos as

reference. Error bars are mean $\pm$ s.e.m. **g**, Percentage of SVs located in different regions of annotated protein-coding genes ($n = 12$). **h**, Heatmap of IBS values based on SVs among 12 genomes. **i**, The proportion of TE-related SVs and non-TE SVs in tetraploid wheat ($n = 11$). Error bars are mean $\pm$ s.e.m. **j**, Number of core, softcore, dispensable and private genes that overlap with SVs ($n = 12$). Statistical significance was determined by one-way analysis of variance followed by Fisher's least-significant difference multiple-comparisons tests. Center line, median; box lower and upper edges, 25% and 75% quartiles, respectively; whiskers, 1.5 $\times$ IQR; colored dots, outliers (**e**, **g** and **j**). DTW, domesticated tetraploid wheat; FTW, free-threshing tetraploid wheat; WEW, wild emmer wheat; CDS, coding sequence.

dispensable and private genic regions (Fig. 3j), demonstrating that SVs confer gene PAVs across different accessions. The extensive sequence diversity and SV landscape in tetraploid wheat preclude accurate variant detection using a single reference genome.

Population structure and subspecies introgression

By integrating the linear reference-genome sequences of Kronos and all SVs identified from the 12 tetraploid wheat genomes, including the published Svevo genome¹³, a comprehensive graph-based



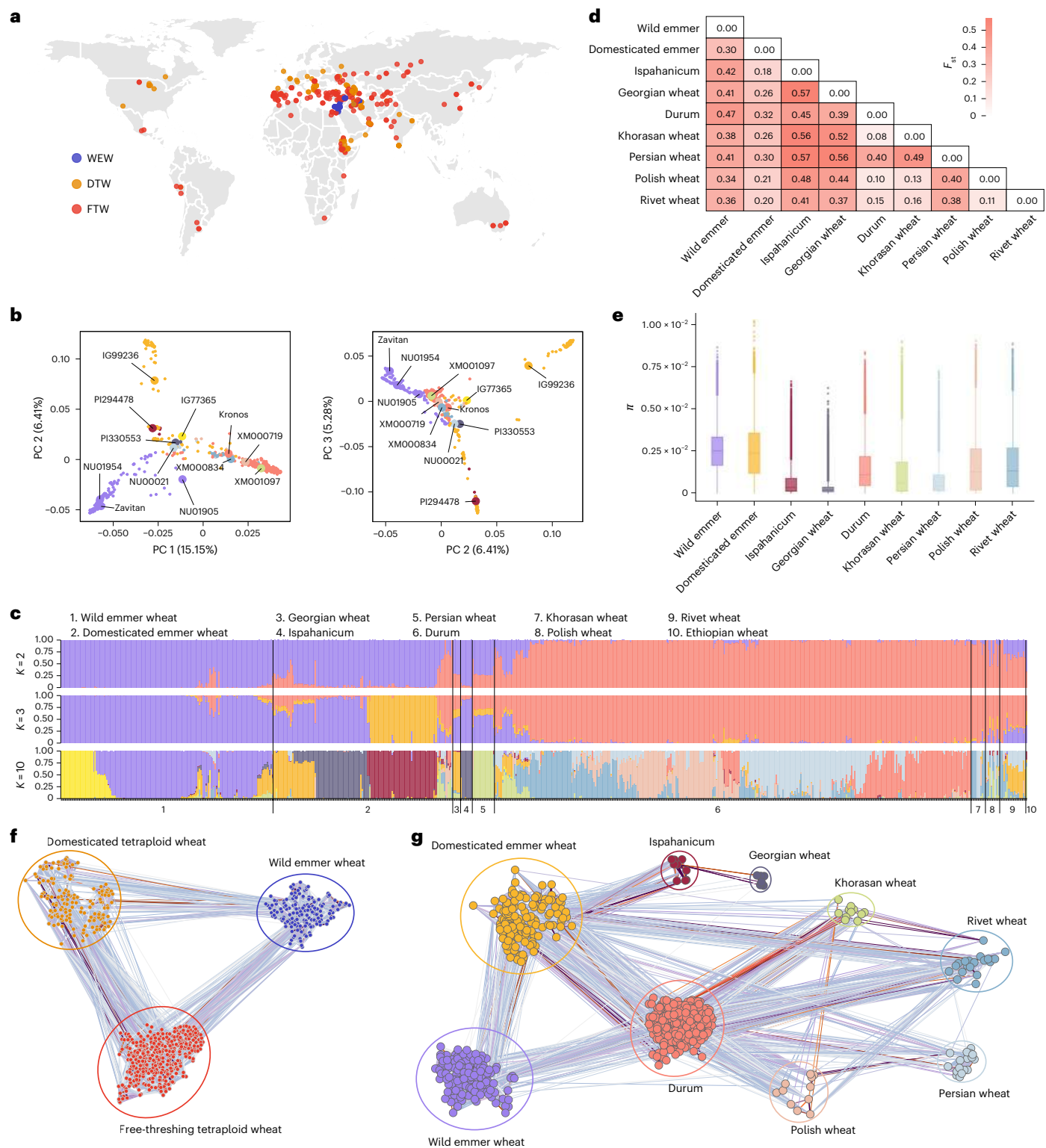


Fig. 4 | Population structure and evolutionary analysis. a, Geographic distribution of 736 sequenced accessions. **b, c**, Principal component (PC) analysis (b) and ancestry coefficient analysis (c) of 736 sequenced accessions. The optimal K value is 10. In **b**, the de novo assembled accessions are highlighted. **d**, Population divergence (F_{st}) among different tetraploid wheat subspecies. **e**, Nucleotide diversity of nine subspecies ($n = 736$). Center line, median; box lower and upper edges, 25% and 75% quartiles, respectively; whiskers, $1.5 \times IQR$. **f, g**, Evolution networks for three major subpopulations (f) and nine subspecies (g) of tetraploid wheat. Clustered phylogenetic consensus genotype networks

of 1,000 maximum-likelihood tree topologies were inferred from a repeated random-haplotype sampling procedure including heterozygous sites. Nodes represent individual genotypes and are color-coded by subspecies. Edges represent minimal evolutionary distances in the repeated random-haplotype sampling trees deduced by the minimum spanning tree algorithm and are color-coded based on distance size. Lines with edge weights lower than the average were removed. The map in **a** was plotted using the R package rworldmap (GPL v.2 license, <https://github.com/AndySouth/rworldmap>), with base map data from Natural Earth (<http://www.naturalearthdata.com>).

pangenome for tetraploid wheat was constructed. Further, we constructed a genetic-variation map by deep whole-genome sequencing of 736 accessions (621 were generated in this study and 115 were publicly available⁸), resulting in 97.11 million single nucleotide polymorphisms (SNPs), 4.36 million small indels (indels, length <50 bp) and 0.59 million SVs (length ≥50 bp) (Fig. 4a and Supplementary Tables 12 and 13), dramatically increasing the availability of tetraploid wheat genomic resources. Phylogenetic analysis suggests that these 736 accessions could be divided into three major subpopulations, namely wild emmer wheat, domesticated tetraploid wheat and free-threshing tetraploid wheat (Supplementary Fig. 20). Principal component analysis and ancestry coefficient analysis also supported the results of three major subpopulations (Fig. 4b,c).

We observed a high degree of differentiation between Ispahanicum (*Triticum ispahanicum*), Georgian (*Triticum karamyschevi*), Persian (*T. carthlicum*) tetraploid wheat and other subspecies, suggesting geographic isolation or ecological adaptation (Fig. 4d). Analysis of nucleotide diversity (π) showed that wild emmer wheat and domesticated emmer wheat (*T. dicoccon*) had higher genetic diversity than the other subspecies; the Ispahanicum and Georgian wheat had the lowest diversity (Fig. 4e). We further investigated the evolution and introgression of those subspecies via an evolution network, with the resulting topology supporting a two-step evolution process for tetraploid wheat (Fig. 4f), from wild emmer wheat to domesticated and free-threshing tetraploid wheat. There are fewer edges for Ispahanicum and Georgian accessions in the network, suggesting less potential introgression from them to free-threshing tetraploid wheats and that they made a smaller contribution to the improvement of cultivated tetraploid wheat (Fig. 4g and Supplementary Note 4). Gene-flow analysis suggested more potential introgression regions arising from wild emmer wheat moved into Persian wheat than other free-threshing tetraploid wheats (Supplementary Note 5 and Supplementary Fig. 21). Except for Persian wheat, the proportion of genome introgression among free-threshing tetraploid wheats was larger than in other types of tetraploid wheats, with an average of 0.36 (Supplementary Table 14). Introgression introduces new genetic variation and contributes to the genomic diversity (Supplementary Figs. 22 and 23a). Genes within the inferred introgression regions showed enrichment in certain pathways, including response to nitrate (GO:0010167) and plant-type cell wall organization (GO:0009664) (Supplementary Table 15). The *GW2* gene is a key determinant of grain weight in rice and bread wheat^{36–38}. *TrdurKRO6BOIG046970*, an ortholog of *TaGW2-B* from bread wheat, is in the putative introgression region ($f_d \approx 0.44$) on chromosome 6B that originated from wild emmer wheat and moved into domesticated emmer wheat (Supplementary Fig. 23b). These results highlight the potential of introgression for improving agricultural traits.

Genetic variation underlying domestication and improvement

To detect putative selection footprints in tetraploid wheat genomes, we compared genetic diversity to identify bottlenecks during

domestication (wild emmer versus domesticated tetraploid wheat) and improvement (domesticated versus free-threshing tetraploid wheat) (Supplementary Fig. 24). Wild emmer wheat exhibited the highest genetic diversity ($\pi = 2.54 \times 10^{-3}$), followed by domesticated tetraploid wheat ($\pi = 2.44 \times 10^{-3}$) and free-threshing tetraploid wheat ($\pi = 1.62 \times 10^{-3}$), indicating a sharp decline in genetic diversity under artificial selection. From a combined SNP- and indel-based selective-sweep analysis, we identified 0.93-Gb and 1.17-Gb selected regions corresponding to domestication and improvement processes, which contained 4,675 and 7,210 genes, respectively (Fig. 5a and Supplementary Table 16). Linkage-disequilibrium analysis revealed that haplotype decay for variants including SVs, SNPs and indels occurred more rapidly than for SNPs alone, suggesting that SVs harbor substantial hidden genetic variation (Supplementary Fig. 25). We compared SV frequencies between wild and landrace accessions to identify putative selected SVs during domestication (Fig. 5b) and between landrace and cultivars for possible selected SVs during improvement. This identified 157,468 SVs associated with 15,642 genes and 151,898 SVs associated with 15,001 genes (Fig. 5b and Supplementary Table 17) exhibiting selection signals during domestication and improvement, of which 133,607 SVs were subjected to selection during both domestication and improvement. Several well-known domesticated genes in the Triticeae tribe were also found, such as *Rht-B1*, *GW8-B1* and *Yr15*^{39–41}. Only 7.48% and 11.97% of selection SVs, during domestication and improvement, respectively, were shared by SNP- and indel-based selection regions (Fig. 5c). A total of 25,379 non-redundant selected genes were identified, of which 2,204 were shared between SNP/indel and SV analysis (Fig. 5d), suggesting the additional power of SVs for selection footprint identification. Enrichment analysis showed selected genes were mainly involved in biological processes, including response to heat (GO:0009408), carbohydrate metabolic process (GO:0005975), positive regulation of long-day photoperiodism and flowering (GO:0048578) and carbon utilization (GO:0015976) (Supplementary Table 18).

Non-brittle rachis phenotype of wheat is a key trait for domestication. Mutations in *Btr1-A* and *Btr1-B* are chiefly responsible for the non-brittle rachis phenotype in tetraploid wheat¹². Several loss-of-function mutations of *Btr1-A* have been observed in domesticated einkorn⁸, barley⁴² and tetraploid wheat¹². Consistent with ref. 12, we observed the hallmark 2-bp deletion introducing a frame shift in the *Btr1-A* coding region of cultivated tetraploids (Fig. 5e). The Persian domesticated tetraploid accession NU00021, with the non-brittle-rachis phenotype, lacks this translation-termination mutation, but has a 5,043-bp LTR/Copia-type TE insertion in the coding region based on the tetraploid wheat pangenome, which likely resulted in loss of its function (Fig. 5f). We also found that both LTRs within this element were identical, hinting that the LTR insertion likely occurred in a recent⁴³ phylogenetic branch of the tetraploid wheat population. Analysis of *Btr1-A* haplotypes across the 736 accessions indicated all sampled Persian accessions contain this LTR insertion (Fig. 5g). These observations suggest an independent loss-of-function led to

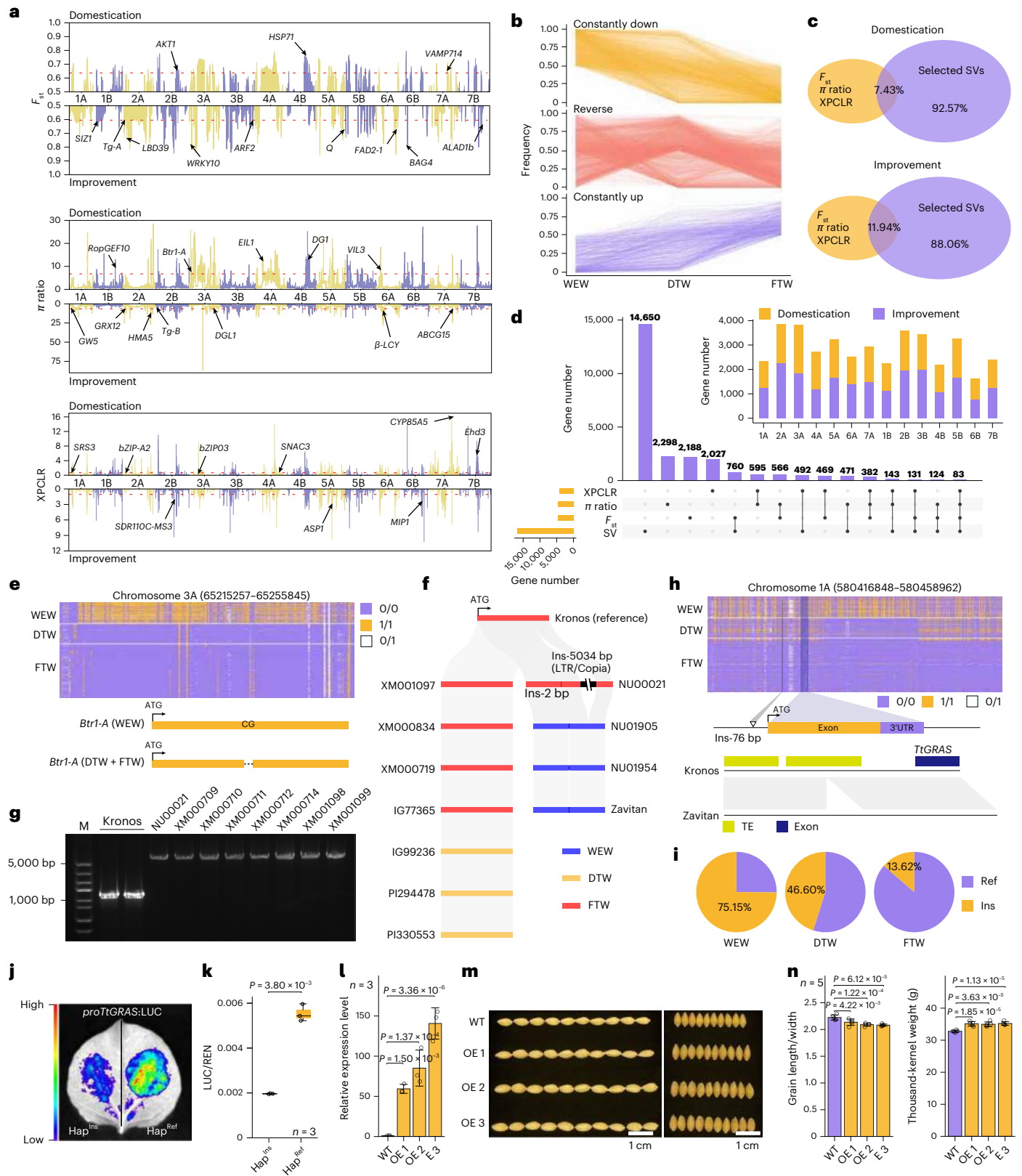
Fig. 5 | Structural variants reveal selection footprints and multiple alleles of *Btr1-A*. **a**, F_{ST} , nucleotide diversity ratio (domestication, π_{WEW}/π_{DTW} ; improvement, π_{DTW}/π_{FTW}) and cross-population composite-likelihood ratio approach (XPCLR) analyses of selection. Dashed lines indicate genome-wide thresholds for selection signals. The upper section of each plot represents the domestication analysis; the lower section denotes the improvement analysis. **b**, Frequency pattern of selection-related SVs. The orange and purple lines represent the gradual decrease and increase in SV frequency during the evolutionary process, respectively. **c**, Overlap of selection signals identified via SV-based, SNP-based and indel-based methods. **d**, Overlap of selected genes across four different methods. **e**, Variation heatmap within 20-kb upstream and downstream of *Btr1-A*. **f**, *Btr1-A* synteny across 12 genomes, wherein *Btr1-A* has a single coding region. **g**, PCR verification of the 5,034-bp TE insertion in the *Btr1-A* coding region found in eight Persian wheats. **h**, Variation heatmap within 20-kb upstream and downstream of *TtGRAS-A* UTR, untranslated region. **i**, Allele frequency

of the selected SV (at position 580,435,572 on chromosome 1A) upstream of *TtGRAS-A* among three subpopulations. Ref, reference (Kronos); Ins, insertion. **j,k**, Chemiluminescence (j) and LUC/REN (k) for dual-reporter assay of Hap^{Ref} and Hap^{Ins} *TtGRAS-A* promoter variants ($n = 3$) transiently expressed in *N. benthamiana* leaves. Statistical significance was determined by two-sided *t*-tests. Center line, median; box lower and upper edges, 25% and 75% quartiles, respectively; whiskers, 1.5 × IQR; colored dots, outliers. **l**, *TtGRAS-A* expression in transgenic wheat lines cv. Fielder with heterologous *TtGRAS-A* expression with three independent biological replicates. Error bars are mean ± s.e.m. **m,n**, Grain phenotypes (**m**) and quantification (**n**) of grain length/width and thousand-kernel weight for transgenic wheat lines heterologously expressing *TtGRAS-A* in the cv. Fielder wild-type background (WT) with five independent biological replicates. Error bars are mean ± s.e.m. Statistical significance was determined by one-way analysis of variance followed by Fisher's least-significant difference multiple-comparisons tests (**l,n**).

convergent evolution of the non-brittle-rachis phenotype⁴⁴. The gene flow and evolution networks suggest that domesticated tetraploid wheat may have multiple origins and Persian wheat may have undergone a different domestication path and history.

In addition, we identified a 76-bp SV-specific locus underlying selection, which is located 1.28-kb upstream of the translational start

codon of a *GRAS* gene, assigned *TtGRAS-A* (*TrdurKRO1A01G076930*). The frequency of the SV-insertion allele decreases from wild (0.75) to domesticated (0.47) and free-threshing tetraploid wheat (0.14) (Fig. 5h,i). This SV on chromosome 1A is statistically significantly associated with thousand-kernel weight and the ratio of grain length/grain width (Supplementary Fig. 26). To examine the regulatory role of this



SV, a transient dual-luciferase reporter assay in *Nicotiana benthamiana* leaves suggested the *TtGRAS-A* SV-deletion allele has higher expression than the SV-insertion allele (Fig. 5j,k). Heterologous expression of *TtGRAS-A* in hexaploid wheat cv. Fielder from the strong, constitutive *Ubiquitin* promoter increased grain size, thousand-kernel weight and effective tiller number per plant (Fig. 5l–n and Supplementary Fig. 27). These results indicate that the long deletion upstream of *TtGRAS-A* may enhance its expression and thereby improve yield-related traits.

Genetic dissection of 32 traits

We phenotyped 373 resequenced tetraploid wheat accessions for 32 agronomically important traits covering key breeding targets: 7 plant-architecture, 8 grain-related and 17 abiotic stress-related traits—to identify variants associated with them (Supplementary Tables 19 and 20). Heritability estimates indicated that SVs based on the graph-based pangenome can explain additional phenotypic variance (Supplementary Fig. 28 and Supplementary Table 21). To dissect the genetic components underlying these traits, we performed SV-based and SNP + indel-based GWAS. A total of 287 loci were associated with 32 traits. These loci contain 783 unique protein-coding genes (Supplementary Fig. 29 and Supplementary Table 22). Of these, 34 SNP + indel-based loci were shared with SV-based approaches and 44 loci were specifically identified by SVs. For instance, a 82-bp SV in the intron of *TrdurKRO3A01G024420* was associated with grain length/width, of which the rice ortholog *RDR4* regulates spike length and grain plumpness⁴⁵ (Supplementary Fig. 30a). Additionally, *TrdurKRO2A01G031640*, an ortholog of rice *SNAC3*⁴⁶, is located 99.3-kb downstream of an indel locus associated with root length under simulated drought treatment (Supplementary Fig. 30b).

HAT14-B^{Hap1} enhances spikelet numbers and grain size

Spikelet number is a critical agronomic trait that is closely associated with wheat grain yield. A 15-bp indel ($P = 6.17 \times 10^{-10}$) on chromosome 1B associated with spikelet number was found in exon 4 of *TrdurKRO1B01G083560* (Fig. 6a). The gene is an ortholog of the *Arabidopsis* homeodomain-leucine zipper (HD-ZIP) transcription factor *HAT14* (Supplementary Fig. 31) and herein designated as *HAT14-B*. This gene has two main haplotypes in tetraploid wheat, *HAT14-B^{Hap1}* and *HAT14-B^{Hap2}* (Fig. 6b and Supplementary Fig. 32). Haplotype-frequency analysis showed that *HAT14-B^{Hap1}* is present in 5.56% (6 of 163) of wild emmer wheat, 11.11% (12 of 147) of domesticated tetraploid wheat and 61.61% (244 of 426) of free-threshing tetraploid wheat accessions after removing the heterozygous sites (Fig. 6c,d), indicating a strong selection footprint from hulled tetraploid wheat to free-threshing tetraploid wheat (Fisher's exact test, $P < 2.2 \times 10^{-16}$). *HAT14-B^{Hap1}* accessions have more spikelet number, higher thousand-kernel weight, higher grain width but lower grain length compared to *HAT14-B^{Hap2}* (Fig. 6e).

In addition, *HAT14-B^{Hap1}* expression in spikes is statistically significantly higher than *HAT14-B^{Hap2}* (Fig. 6e). Variation in the *HAT14-B* coding region did not obviously alter the nuclear localization of the corresponding protein products when fused to GFP (Fig. 6f). HD-ZIP transcription factors bind as dimers to palindromic DNA sequences containing the AATNATT motif (AT-containing motif)⁴⁷. *HAT14-B^{Hap2}* DNA-binding ability was attenuated compared to that of *HAT14-B^{Hap1}* in electrophoretic mobility shift assay (EMSA) experiments (Fig. 6g). Dual-luciferase assays showed relatively high activity from an approximately 2-kb *HAT14-B^{Hap1}* promoter fragment compared to the *HAT14-B^{Hap2}* promoter fragment (Fig. 6h). These results indicated that *Hap1* is the elite allele of *HAT14-B*.

To explore the function of the elite *HAT14-B^{Hap1}* allele in spike development, we heterologously expressed its coding region in hexaploid wheat cv. Fielder (Fig. 6i). Transgenic lines had statistically significantly greater spike length, spikelet number, thousand-kernel weight, grain length and grain width compared to Fielder controls (Fig. 6j,k), affirming that the elite *HAT14-B^{Hap1}* allele is a robust candidate for regulating spikelet number and grain size. Furthermore, RNA-mediated interference (RNAi) lines against *HAT14-B^{Hap1}* had shorter spike length, fewer spikelet number, reduced thousand-grain weight and smaller grains compared to Fielder controls (Fig. 6l,m). However, the *HAT14-B^{Hap1}* transgenic and RNAi lines did not exhibit a significant difference in effective tiller number (Supplementary Fig. 33). We demonstrate that *HAT14-B^{Hap1}* enhances spikelet number and grain size in wheat. Those discoveries, enabled by the newly generated tetraploid wheat pangenome, underscore its utility for gene discovery and crop improvement.

Discussion

Tetraploid wheat (*T. turgidum* L., BBAA), the donor of B and A subgenomes of hexaploid bread wheat, possesses abundant genetic and germplasm resources⁴⁸. Tetraploid wheat serves as a vital gene pool with abundant untapped beneficial alleles and loci for improving bread wheat agronomic traits. Nevertheless, efficient genomic molecular breeding and gene editing are limited by insufficient comprehensive in silico resources. Here we performed de novo assembly of 12 high-quality genomes covering all 10 known tetraploid wheat subspecies to support the mining of novel alleles and SVs.

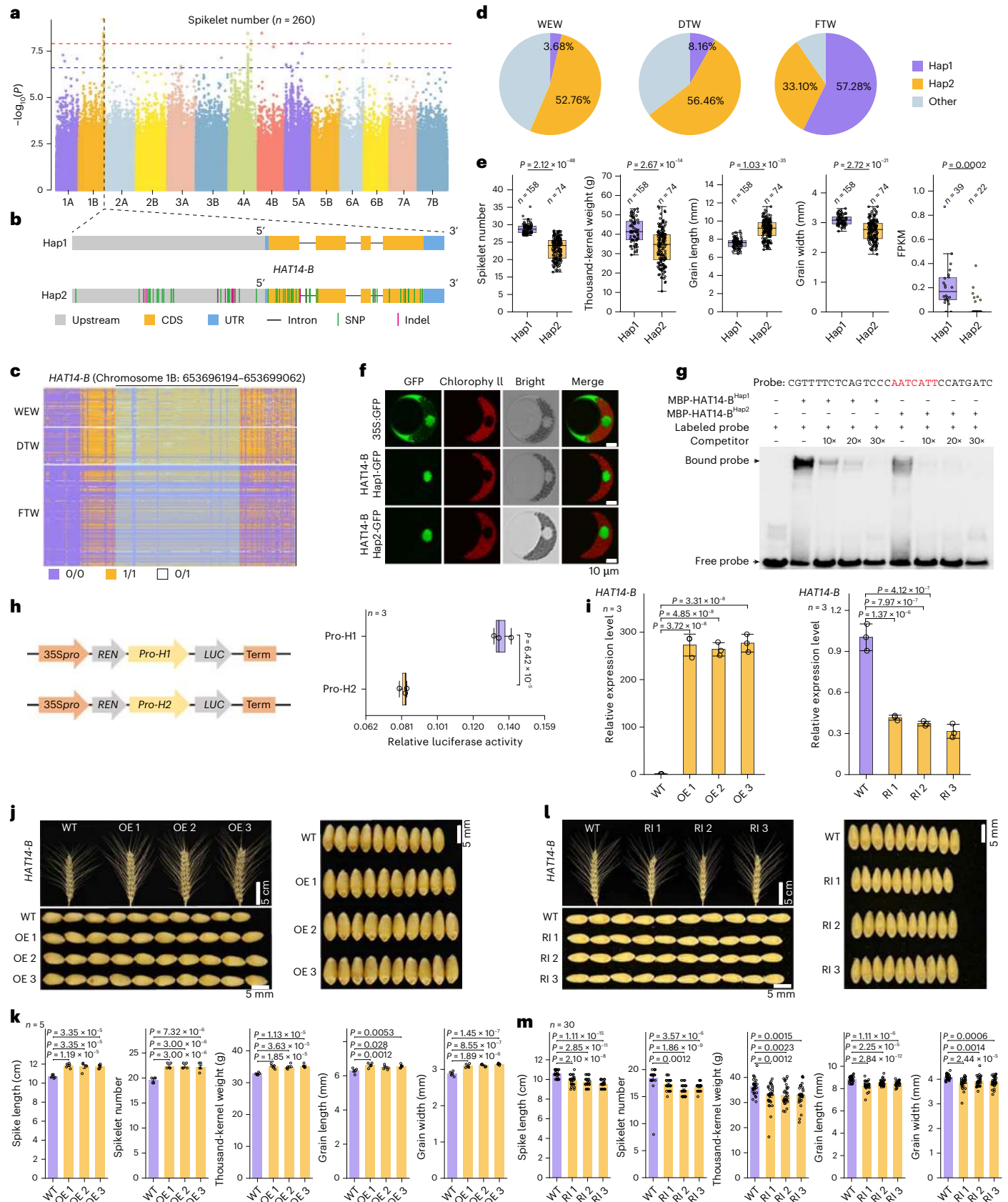
Pangenome analysis revealed substantial genomic diversity in tetraploid wheat subspecies. Phylogenetic analysis using orthologous genes clearly differentiated wild emmer, domesticated tetraploid and free-threshing wheat into separate clades and clarified the evolutionary trajectories of the ten subspecies. Disease resistance genes were enriched in contracted gene families during the transition from domesticated to free-threshing tetraploid wheat. We also identified numerous tetraploid wheat-specific NLR genes, offering valuable new genetic resources for disease-resistance breeding⁴⁹.

Fig. 6 | The HD-ZIP transcription factor *HAT14-B* positively regulates spikelet number and grain size. **a**, Indel-based GWAS of 260 resequenced accessions with spikelet number phenotype. The horizontal lines indicate the two-sided Bonferroni-corrected genome-wide statistical significance threshold ($P = 2.37 \times 10^{-7}$ ($\alpha = 1$) and $P = 1.83 \times 10^{-8}$ ($\alpha = 0.05$)). **b**, Gene structure and haplotype analysis of *HAT14-B*. **c**, Variation heatmap from 2-kb upstream to 2-kb downstream of *HAT14-B*. **d**, Proportions of WEW, DTW and FTW accessions with the two *HAT14-B* haplotypes. **e**, Box plots for spikelet number, thousand-kernel weight, grain length, grain width and *HAT14-B* expression for accession with *HAT14-B^{Hap1}* or *HAT14-B^{Hap2}*. Statistical significance was determined by two-sided *t*-tests. FPKM, fragments per kilobase of transcript per million mapped reads. **f**, Subcellular localization of *HAT14-B^{Hap1}* and *HAT14-B^{Hap2}* fused to GFP in transiently transgenic wheat protoplasts with twice repeats. **g,h**, EMSA analysis of probes bearing the AT-containing motif (**g**) and promoter activity via dual-reporter assays (**h**) of *HAT14-B^{Hap1}* and *HAT14-B^{Hap2}*. Values are mean $\pm$ s.e.m. from three independent biological replicates and EMSAs were repeated twice. Statistical significance was determined by two-sided *t*-tests. **i**, *HAT14-B* expression in transgenic wheat lines with heterologous *HAT14-B* expression (left)

and RNAi knock-down lines (right), all in the wild-type cv. Fielder background. Values are mean $\pm$ s.e.m. from three independent biological replicates. Statistical significance was determined by one-way analysis of variance followed by Fisher's least-significant difference multiple-comparisons tests. **j,k**, Representative spike and grain phenotypes (**j**) and quantification (**k**) of spike length, spikelet number, thousand-kernel weight, grain length and grain width for transgenic wheat lines heterologously expressing *HAT14-B^{Hap1}* in the wild-type cv. Fielder background. Values represent the mean $\pm$ s.e.m. from five independent biological replicates. **l,m**, Representative spike and grain phenotypes (**l**) and quantification (**m**) of spike length, spikelet number, thousand-kernel weight, grain length and grain width for transgenic *HAT14* RNAi lines in the wild-type cv. Fielder background. Values represent the mean $\pm$ s.e.m. from 30 independent biological replicates. Statistical significance was determined by one-way analysis of variance followed by Fisher's least-significant difference multiple-comparisons test (**k,m**). Center line, median; box lower and upper edges, 25% and 75% quartiles, respectively; whiskers, 1.5 $\times$ IQR; colored dots, outliers. OE, heterologous-expression lines; RI, RNAi lines; WT, wild type.

Extensive SVs, SNPs and indels further shape the high genomic diversity of tetraploid wheat. Notably, pairwise accessions exhibit an average sequence divergence of 12.58%, reflecting its remarkable genetic variation. Leveraging these high-quality genomes, we

constructed a graph-based pangenome. This study reports a comprehensive pangenome of tetraploid wheat (*T. turgidum* L., BBAA) covering all tetraploid wheat subspecies and dissecting their broad genetic diversity. This dataset represents the most complete atlas of



tetraploid wheat genomic diversity to date, providing key resources for dissecting evolutionary mechanisms and enriching the genetic reservoir for bread wheat improvement.

Genomic signatures of domestication and improvement frequently offer practical insights for molecular breeding^{50,51}. Population evolution and dynamics of tetraploid wheat have been examined, but the evolutionary relationships among different subspecies remain unclear^{7,8,18,52}. We characterized the evolutionary history of ten tetraploid wheat subspecies, finding that Ispahanicum and Georgian subspecies bear high genetic differentiation and low introgression with other subspecies, likely due to geographic isolation. Persian wheat is also markedly divergent from other free-threshing tetraploids and harbors a loss-of-function *Btr1-A* allele without the canonical 2-bp deletion¹⁸, while having a 5,034-bp LTR/Copia insertion in the protein-coding region. A recent study indicates that independent loss-of-function alleles likely emerged in wild emmer populations⁴⁴, highlighting the complexity of crop evolution. These findings underscore the potential of pangenomic approaches to reveal novel domestication variants and provide new targets for accelerated crop improvement and de novo domestication⁵³. The development of new wheat varieties is urgently required to adapt to climate change and future planting. Pangenome-driven strategies facilitate the mining of untapped genetic variations for crop breeding. Unlike single-genome references, pangenomes enable comprehensive dissection of genetic mechanisms governing phenotypic diversity. Similar to earlier findings, SVs conferred extra heritability beyond SNPs and indels, highlighting the strengths of pangenome-supported SV-GWAS^{54,55}.

Using the tetraploid wheat pangenome and GWAS, we identified the pleiotropic gene *HAT14-B*, which provides an important resource for breeding of high-yielding wheat varieties. The tetraploid wheat pangenome complements existing hexaploid wheat resources and will accelerate the identification of key trait genes as well as facilitate the breeding of improved wheat varieties by tapping into the rich genetic diversity of wheat ancestors.

Overall, this study provides a foundational genomic resource that will accelerate comparative genomics and breeding in wheat and pave the way for the effective deployment of tetraploid wheat diversity in improving wheat.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41588-026-02678-9>.

References

- Appels, R. et al. (eds) *The Wheat Genome* (Springer, 2023).
- Nesbitt, M. in *Wheat Taxonomy: the legacy of John Percival* (eds Caligari, P.D.S. & Brandham, P. E.) 37–59 (Academic, 2001).
- Mohammadi, R. & Haghparast, R. Durum wheat: production, nutritional value and economic importance. *Cereal Biotechnol. Biochem.* **1**, 414–445 (2022).
- De Vita, P. & Taranto, F. in *Advances in Plant Breeding Strategies: Cereals* (eds Al-Khayri, J. M. et al.) 471–524 (Springer, 2019).
- Yang, F. et al. Improvement and re-evolution of tetraploid wheat for global environmental challenge and diversity consumption demand. *Int. J. Mol. Sci.* **23**, 2206 (2022).
- Salamini, F., Ozkan, H., Brandolini, A., Schäfer-Pregl, R. & Martin, W. Genetics and geography of wild cereal domestication in the near east. *Nat. Rev. Genet.* **3**, 429–441 (2002).
- Levy, A. A. & Feldman, M. Evolution and origin of bread wheat. *Plant Cell* **34**, 2549–2567 (2022).
- Zhou, Y. et al. *Triticum* population sequencing provides insights into wheat adaptation. *Nat. Genet.* **52**, 1412–1422 (2020).
- Shen, T. et al. The wheat *Sr8155B1* gene encodes a typical NLR protein that confers resistance to the Ug99 stem rust race group. *Plant Commun.* **6**, 101296 (2025).
- Uauy, C., Brevis, J. C. & Dubcovsky, J. The high grain protein content gene *Gpc-B1* accelerates senescence and has pleiotropic effects on protein content in wheat. *J. Exp. Bot.* **57**, 2785–2794 (2006).
- Klymiuk, V. et al. Coordinated function of paired NLRs confers Yr84-mediated stripe rust resistance in wheat. *Nat. Genet.* **57**, 1535–1542 (2025).
- Avni, R. et al. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* **357**, 93–97 (2017).
- Maccaferri, M. et al. Durum wheat genome highlights past domestication signatures and future improvement targets. *Nat. Genet.* **51**, 885–895 (2019).
- Chen, M. et al. Chromosome-level genome assembly of the Durum wheat cultivar Langdon. *Sci. Data* **12**, 1372 (2025).
- Gent, J. I., Wang, K., Jiang, J. & Dawe, R. K. Stable patterns of CENH3 occupancy through maize lineages containing genetically similar centromeres. *Genetics* **200**, 1105–1116 (2015).
- Chen, J. et al. Pangenome analysis reveals genomic variations associated with domestication traits in broomcorn millet. *Nat. Genet.* **55**, 2243–2254 (2023).
- Zhang, Y. et al. Telomere-to-telomere *Citrullus* super-pangenome provides direction for watermelon breeding. *Nat. Genet.* **56**, 1750–1761 (2024).
- Zhao, X. et al. Population genomics unravels the Holocene history of bread wheat and its relatives. *Nat. Plants* **9**, 403–419 (2023).
- Zohary, D., Hopf, M. & Weiss, E. *Domestication of Plants in the Old World: The Origin and Spread of Domesticated Plants in Southwest Asia, Europe, and the Mediterranean Basin* (Oxford Univ. Press, 2012).
- Nesbitt, M. in *Hulled Wheat* (ed. Padulosi, S.) 41–100 (IPGRI, 1996).
- Peng, J., Sun, D. & Nevo, E. Wild emmer wheat, *Triticum dicoccoides*, occupies a pivotal position in wheat domestication process. *Aust. J. Crop Sci.* **5**, 1127–1143 (2011).
- Luo, M.-C. et al. The structure of wild and domesticated emmer wheat populations, gene flow between them, and the site of emmer domestication. *Züchter Genet. Breed. Res.* **114**, 947–959 (2007).
- Feldman, M. & Kislev, M. E. Domestication of emmer wheat and evolution of free-threshing tetraploid wheat. *Isr. J. Plant Sci.* **55**, 207–221 (2007).
- Walkowiak, S. et al. Multiple wheat genomes reveal global variation in modern breeding. *Nature* **588**, 277–283 (2020).
- Jiao, C. et al. Pan-genome bridges wheat structural variations with habitat and breeding. *Nature* **637**, 384–393 (2025).
- Ning, W. et al. The pan-NLRome analysis based on 23 genomes reveals the diversity of NLRs in *Brassica napus*. *Mol. Breed.* **44**, 2 (2024).
- Van de Weyer, A.-L. et al. A species-wide inventory of NLR genes and alleles in *Arabidopsis thaliana*. *Cell* **178**, 1260–1272 (2019).
- Li, S.-X., Liu, Y., Zhang, Y.-M., Chen, J.-Q. & Shao, Z.-Q. Convergent reduction of immune receptor repertoires during plant adaptation to diverse special lifestyles and habitats. *Nat. Plants* **11**, 248–262 (2025).
- Guo, B.-C. et al. Deciphering plant NLR genomic evolution: synteny-informed classification unveils insights into TNL gene loss. *Mol. Biol. Evol.* **42**, msaf015 (2025).
- Zhu, K. et al. An atypical NLR pair *TdCNL1/TdCNL5* from wild emmer confers powdery mildew resistance in wheat. *Nat. Genet.* **57**, 1553–1562 (2025).
- Hu, Y. et al. A head-to-head NLR gene pair from wild emmer confers stripe rust resistance in wheat. *Nat. Genet.* **57**, 1543–1552 (2025).

32. Dvorak, J. et al. Reassessment of the evolution of wheat chromosomes 4A, 5A, and 7B. *Züchter Genet. Breed. Res.* **131**, 2451–2462 (2018).
33. Devos, K. M., Dubcovsky, J., Dvořák, J., Chinoy, C. N. & Gale, M. D. Structural evolution of wheat chromosomes 4A, 5A, and 7B and its impact on recombination. *Züchter Genet. Breed. Res.* **91**, 282–288 (1995).
34. Badaeva, E. D. et al. Evolution of *Triticum aethiopicum* Jakubz. from the position of chromosome analysis. *Russ. J. Genet.* **54**, 629–642 (2018).
35. Mun, S. et al. A study of transposable element-associated structural variations (TASVs) using a de novo-assembled Korean genome. *Exp. Mol. Med.* **53**, 615–630 (2021).
36. Zhang, Y. et al. Analysis of the functions of *TaGW2* homoeologs in wheat grain weight and protein content traits. *Plant J.* **94**, 857–866 (2018).
37. Li, S. et al. The E3 ligase *TaGW2* mediates transcription factor *TaARR12* degradation to promote drought resistance in wheat. *Plant Cell* **36**, 605–625 (2024).
38. Song, X.-J., Huang, W., Shi, M., Zhu, M.-Z. & Lin, H.-X. A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. *Nat. Genet.* **39**, 623–630 (2007).
39. Ma, L. et al. Diversity and sub-functionalization of *TaGW8* homoeologs hold potential for genetic yield improvement in wheat. *Crop J.* **7**, 830–844 (2019).
40. Hedden, P. The genes of the Green Revolution. *Trends Genet.* **19**, 5–9 (2003).
41. Klymiuk, V. et al. Cloning of the wheat *Yr15* resistance gene sheds light on the plant tandem kinase-pseudokinase family. *Nat. Commun.* **9**, 3735 (2018).
42. Pourkheirandish, M. et al. Evolution of the grain dispersal system in barley. *Cell* **162**, 527–539 (2015).
43. Jedlicka, P., Lexa, M. & Kejnovsky, E. What can long terminal repeats tell us about the age of LTR retrotransposons, gene conversion and ectopic recombination? *Front. Plant Sci.* **11**, 644 (2020).
44. Cavalet-Giorsa, E., Wicker, T. & Krattinger, S. G. Tracing the origin of non-brittle rachis alleles in wheat. Preprint at *bioRxiv* <https://doi.org/10.64898/2026.02.26.708176> (2026).
45. Jha, V. et al. Essential role of γ -clade RNA-dependent RNA polymerases in rice development and yield-related traits is linked to their atypical polymerase activities regulating specific genomic regions. *New Phytol.* **232**, 1674–1691 (2021).
46. Fang, Y. et al. A stress-responsive NAC transcription factor SNAC3 confers heat and drought tolerance through modulation of reactive oxygen species in rice. *J. Exp. Bot.* **66**, 6803–6817 (2015).
47. Franco-Zorrilla, J. M. et al. DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proc. Natl Acad. Sci. USA* **111**, 2367–2372 (2014).
48. Feldman, M. & Levy, A. A. *Wheat Evolution and Domestication* (Springer, 2023).
49. Lu, P. et al. A wheat tandem kinase and NLR pair confers resistance to multiple fungal pathogens. *Science* **387**, 1418–1424 (2025).
50. Huang, X., Huang, S., Han, B. & Li, J. The integrated genomics of crop domestication and breeding. *Cell* **185**, 2828–2839 (2022).
51. Doebley, J. F., Gaut, B. S. & Smith, B. D. The molecular genetics of crop domestication. *Cell* **127**, 1309–1321 (2006).
52. Sertse, D. et al. Genome scans capture key adaptation and historical hybridization signatures in tetraploid wheat. *Plant Genome* **18**, e20410 (2025).
53. Schreiber, M., Jayakodi, M., Stein, N. & Mascher, M. Plant pangenomes for crop improvement, biodiversity and evolution. *Nat. Rev. Genet.* **25**, 563–577 (2024).
54. Zhou, Y. et al. Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* **606**, 527–534 (2022).
55. Liu, Z. et al. Grapevine pangenome facilitates trait genetics and genomic breeding. *Nat. Genet.* **56**, 2804–2814 (2024).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026

Jianxin Bian (卞建新)^{1,10}, Guang Yang (杨光)^{1,2,3,10}, Dong Xu (许栋)^{1,2,10}, Yan Zhang (张岩)^{1,2,10}, Yanzhe Jia (贾彦哲)^{2,10}, Guifen Zhang (张贵粉)¹, Zhen Qin (秦震)², Shuangxing Zhang (张双星)², Yiqiao Wang (王艺桥)¹, Mengmeng Jiang (姜蒙蒙)¹, Yan Pan (潘燕)^{2,4}, Bin Chen (陈斌)^{1,2}, Fangyu Liu (刘芳宇)^{1,2}, Yuqing Lu (鲁玉清)⁵, Shuai Ding (丁帅)¹, Jiabo Wang (王嘉博)⁶, Can Yuan (袁灿)¹, Yongming Chen (陈永明)¹⁰, Shoucheng Liu (刘守成)¹⁰, Hang He (何航)¹⁰, Shuai Wang (王帅)^{1,2}, Qidi Zhu (朱启迪)⁸, Shisheng Chen (陈时盛)¹⁰, Qing Sang (桑青)^{1,9}, Xing Wang Deng (邓兴旺)¹⁰, Hude Mao (毛虎德)¹⁰, Xiaojun Nie (聂小军)¹⁰ & Baoxing Song (宋宝兴)^{1,2}

¹State Key Laboratory of Wheat Improvement, Peking University Institute of Advanced Agricultural Sciences, Shandong Laboratory of Advanced Agricultural Sciences at Weifang, Weifang, China. ²State Key Laboratory for Crop Stress Resistance and High-Efficiency Production, College of Agronomy, Northwest A&F University, Yangling, China. ³Tea Research Institute, Chongqing Academy of Agricultural Sciences, Chongqing, China. ⁴Oil Crop Research Institute, Chongqing Academy of Agricultural Sciences, Chongqing, China. ⁵National Key Facility for Crop Gene Resources and Genetic Improvement, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China. ⁶Key Laboratory of Qinghai-Tibetan Plateau Animal Genetic Resource Reservation and Utilization, Sichuan Province and Ministry of Education, Southwest Minzu University, Chengdu, China. ⁷Peking-Tsinghua Center for Life Sciences, School of Life Sciences and School of Advanced Agricultural Sciences, Peking University, Beijing, China. ⁸School of Agriculture, Henan Institute of Science and Technology, State Key Laboratory of High-Efficiency Production of Wheat-Maize Double Cropping, Xinxiang, China. ⁹Innovative Center of Molecular Genetics and Evolution, School of Life Sciences, Guangzhou University, Guangzhou, China. ¹⁰These authors contributed equally: Jianxin Bian, Guang Yang, Dong Xu, Yan Zhang, Yanzhe Jia. ✉ e-mail: shisheng.chen@pku-iaas.edu.cn; qing.sang@pku-iaas.edu.cn; deng@pku.edu.cn; mao_dehu@nwsuaf.edu.cn; small@nwsuaf.edu.cn; baoxing.song@pku-iaas.edu.cn

Methods

Plant material

A total of 622 accessions covering all of the ten subspecies belonging to tetraploid wheat were used in this study, representing a worldwide tetraploid wheat germplasm collection (Supplementary Tables 19 and 20 and Supplementary Note 1). We conducted de novo assembly for 12 accessions aiming to cover 10 subspecies from different areas (Supplementary Tables 1 and 12). Among them, three wild emmer accessions were de novo assembled and they harbor high genomic diversity of the wild subspecies, from different regions of Levant, indicated by the principal component analysis.

Genome assembly

Step 1. Contig construction. We used a hybrid approach for initial contig assembly. For combined HiFi and ONT ultra-long reads (length >100 kb, QV > 10), we used both hifiasm (v.0.22.0-r689)⁵⁶ and Verkko (v.2.0)⁵⁷ assemblers. We also used hifiasm (v.0.22.0-r689) to conduct assembly using ONT data independently. Considering the high computational resource cost of NextDenovo (v.2.5.2)⁵⁸ and the large genome size of Kronos, we implemented a chromosome-by-chromosome assembly strategy. This involved aligning ONT data to 14 chromosomes using minimap2 (v.2.24)⁵⁹ (targeting only HiFi-assembled regions), followed by separate NextDenovo (v.2.5.2)⁵⁸ assemblies for each chromosomal ONT subset. To filter organelle genomes, the chloroplast and mitochondrial genomes of *Triticum urartu*, *Triticum monococcum*, *Triticum zhukovskyi*, *T. turgidum* (cv. Svevo), *Triticum timopheevii* and *Triticum macha* were downloaded from the National Center for Biotechnology Information. We used minimap2 (v.2.24)⁵⁹ with -cx asm20 to align the contigs of each species onto the reference organelle genomes. Contigs that are composed of ≥95% organelle sequence and ≤5% divergence were filtered to obtain nuclear genomes.

Step 2. Hi-C scaffolding. Chromosomal contigs were integrated through hierarchical scaffolding. We first applied YAHS (v.1.2a.1)⁶⁰ for initial scaffolding of different assembly outputs. Subsequently, Rag-Tag (v.2.1.0)⁶¹ was used to refine the NextDenovo (v.2.5.2)⁵⁸ assembly results. To address gaps in the primary framework, we extracted two types of sequences: (1) unaligned ONT contigs; and (2) 2-Mb flanking sequences at scaffold gaps. These components underwent dedicated NextDenovo (v.2.5.2)⁵⁸ assembly to generate supplementary contig sets.

Step 3. Gap filling. The Hi-C scaffold hifiasm-derived hybrid assembly (HiFi + ONT) was designated as the reference framework. We used the genoAli command of AnchorWave (v.1.2.6)⁶² to establish a global collinearity map between an alternative assembly and the reference genome for each chromosome. For gap filling, the interval between each pair of adjacent collinear anchor pairs was systematically evaluated: reference-genome intervals containing gaps while absent from corresponding alternative assembly regions underwent sequence replacement using the contiguous sequence of the alternative assembly. This gap-filling protocol was iteratively applied across four alternative assemblies to maximize gap-closure efficiency.

Step 4. Sequence polishing. Three rounds of polishing were performed using NextPolish (v.1.4.1)⁶³, exclusively using HiFi reads for error correction.

Assembled genome evaluation

To evaluate the completeness of each genome assembly, BUSCO (v.5.4.6)⁶⁴ was used with the OrthoDB database poales_odb10 (poales, 2020-09-10). We evaluated the consensus QV and completeness of genome assemblies on the basis of the *k*-mers spectrum of HiFi reads and Illumina whole-genome sequencing reads, using Merquary (v.1.3)⁶⁵ with default parameters. To evaluate assembly continuity, LAI was calculated for each genome assembly using LTR_retriever (v.2.9.0)⁶⁶.

Repeat annotation

Tandem repeats of all assemblies were identified using Tandem Repeats Finder (TRF) (v.4.09)⁶⁷. We also identified candidate telomeres on scaffolds based on TRF (v.4.09) results using the plant telomeric-repeat motif (CCCTAAA)*n* or (TTAGGG)*n*. TEs were identified by the Extensive de novo TE Annotator (EDTA) (v.2.0.1)⁶⁸. In brief, structure-intact TEs (including LTR-RT, DNA transposon, Helitron) predicted by EDTA (v.2.0.1)⁶⁸ were firstly softmasked.

The unknown TE sequences in the fragmented TE library predicted by EDTA (v.2.0.1)⁶⁸ were classified using the Transposable Element Representation Learner (TERL)⁶⁹. The known and classified fragmented TE library was then used to mask the genome with RepeatMasker (<http://repeatmasker.org/>; v.4.1.1). Finally, we merged all TE annotations, removed redundancy and obtained the final TE annotation and the softmasked genome of each assembly.

Genome annotation

To minimize discrepancies arising from the use of different annotation tools, we used a unified annotation pipeline for all 12 genomes assembled de novo. The annotation integrated evidence from RNA-seq, Iso-seq, homologous protein alignments and ab initio gene prediction.

Stranded RNA-seq reads were aligned to softmasked genome using STAR⁷⁰ (v.2.7.10b; --outSAMattrIHstart 0 --outSAMstrandField intronMotif --alignIntronMin 20 --alignIntronMax 20000) and transcripts were assembled using StringTie⁷¹ (v.3.0.0; -m 150 -t -f 0.3 -rf). Iso-seq data were processed by first mapping redundant isoforms to the softmasked genome using pbmm2 (<https://github.com/PacificBiosciences/pbmm2>; v.1.17.0), followed by collapsing into transcripts using Iso-seq (<https://github.com/PacificBiosciences/IsoSeq>; v.4.3.0). Transcript loci were predicted by GMAP⁷² (v.2024-11-20; -L 10000000). For coding-region prediction, transcript assemblies from RNA-seq and Iso-seq were merged using StringTie⁷¹ (v.3.0.0; -m 150) and passed to TransDecoder (<https://github.com/TransDecoder/TransDecoder>; v.5.7.1), resulting in transcriptome-derived gene annotations.

For protein homology evidence, gene structures from a diverse panel of species, including *Arabidopsis thaliana* (TAIR10), *Sorghum bicolor* (v.3), *Zea mays* (B73 NAM V5), *O. sativa* (IRGSP-1.0.60), *Secale cereale*, *Brachypodium distachyon* (v.3.0), *Aegilops speltoides*, *Hordeum vulgare*, *T. aestivum* (IWGSC v2.1), *T. dicoccoides*, *T. turgidum* (Svevo.v1) and *T. urartu* were projected onto softmasked genomes using miniprot (v.0.12)⁷³.

Ab initio gene predictions were produced with Augustus⁷⁴ (v.3.5.0; --softmasking=on --genemodel=complete --alternatives-from-evidence = on) and GeneMark-ET⁷⁵ (v.4.72.lic). We merged the transcriptome-derived gene annotations and homology evidence using the cuffcompare program from cufflinks⁷⁶ (v.2.2.1) and StringTie⁷¹ (v.3.0.0) and passed the resulting GTF file to TransDecoder (v.5.7.1) to generate training data for Augustus⁷⁴ (v.3.5.0). Augustus⁷⁴ (v.3.5.0) gene prediction was performed using a model specifically trained from the software and three hints files generated by the above-mentioned RNA-seq, Iso-seq and homology-based predictions. GeneMark-ET⁷⁵ (v.4.72.lic) was used with the option -ET and intron coordinates were calculated using the above-mentioned RNA-seq alignments.

All the gene evidence from transcriptomics, protein alignments and ab initio predictions was integrated using EvidenceModeler⁷⁷ (v.2.1.0; --segmentSize 100000 --overlapSize 10000), with weights assigned using the parameters in ref. 78, transcript (merged from RNA-seq and Iso-seq, 10), homology (8), Augustus (5) and GeneMark (2). The resulting annotations were refined through two rounds of PASA (<https://github.com/PASAPipeline/PASAPipeline>; v.2.5.3).

Gene models were classified as high-confidence (HC) or low confidence (LC) genes according to criteria used by the IWGSC (https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_RefSeq_Annotations/v1.0/iwgsc_refseqv1.0_README.pdf) and ref. 79. DIAMOND (v.2.1.8)⁸⁰ was used to compare predicted protein sequences against

three curated datasets: UniMag (38,182 reviewed Magnoliopsida proteins, downloaded from Uniprot (SwissProt), June 2025), UniPoa (2,848,452 Poaceae proteins downloaded from Uniprot (SwissProt and trEMBL), June 2025)⁸¹ and PTREP (<https://trep-db.uzh.ch/index.php>; Rel-19, June 2025). Protein-encoding gene models were considered complete when translational start and stop codons were present. A HC protein sequence is complete with a hit in the UniMag database (HC1), or with hits in UniPoa and not in TREP (HC2). An LC protein sequence is incomplete and has a hit in the UniMag (LC1) or complete gene models with no hits to any of the three databases (LC2) or incomplete gene models with hits in Poaceae protein but no hits in the TE database TREP (LC3) or incomplete gene models with no hits to any of the three databases (LC4). TREP genes have no hits to UniMag proteins but with hits to TREP entries. We also promoted gene models with intron chains supported by Iso-seq to HC (HC3). InterProScan (v.5.73)⁸² was used to predict potential protein domains and perform gene ontology annotations for HC genes of each accession with `-goterms -iprlookup -pa -dp`.

Centromeric region identification

Adapters and low-quality reads in ChIP-seq data for 12 accessions were removed using Trimmomatic (v.0.39)⁸³ with ILLUMINACLIP: adapter. fa:2:30:10 MINLEN:36 SLIDINGWINDOW:4:20. Quality-controlled reads were aligned to each genome using Bowtie2⁸⁴ (v.2.5.1; `--very-sensitive --no-mixed --no-discard -k 10 -X 1000`) and were further filtered by SAMtools⁸⁵ (v.1.10; `-F 1804 -f 2 -q 30`). MACS3 (v.3.0.1)⁸⁶ was used to call CENH3 ChIP-seq peaks using input as a control with `-f BAMPE`. Adjacent CENH3-enriched regions within 1,000 kb were merged and the most enriched regions were determined as potential centromere regions. For visualization, aligned and filtered BAM files were converted to normalized coverage files (bigwig) using Deeptools (v.3.5.5)⁸⁷ with default parameters. The structure of tandem repeat sequences in each centromere region was confirmed using StainedGlass⁸⁸ (v.0.6; `window = 10,000`).

Pangenome analysis

Orthofinder⁸⁹ (v.2.5.5) was used to construct the tetraploid wheat gene families, with `-M msa -S blast -T iqtree`. All gene families were classified into four categories: core (gene families present in all accessions), softcore (gene families present in 12 accessions), dispensable (gene families present in 2–11 accessions) and private (gene families present in one accession). To characterize the changes in the number of pan genes (dispensable and private genes) and core genes, we conducted 500 random samplings for each genome count scenario (ranging from 2 to 13) and plotted the variation curves using ggplot2⁹⁰. The ratio of non-synonymous to synonymous substitutions (K_a/K_s) between each gene in the pangenome was estimated by the Nei–Gojori approach⁹¹ built in the Bioperl Statistical module using barley as an outgroup based on sequence alignments performed with BLAST+ (v.2.16.0)⁹².

On the basis of the genome-annotation file, TE annotation results and pangenome data, the distribution characteristics of TEs in and around (± 2 kb) all protein-coding genes in the pangenome were analyzed. For each gene, the transcriptional start/termination sites (TSS/TTS) were used as the centre to examine the 2-kb upstream and downstream regions. A sliding-window approach with adjustable window size (200 bp) and step size (20 bp) was used to calculate TE density. If a window extended beyond the chromosome boundaries, its TE density was marked as an invalid value. For the gene-body region, a fixed number of 100 windows was used. If the gene length was < 2 kb, the number of valid windows was set to $(\text{gene length})/20$, with the remaining windows marked as invalid. If the gene length was ≥ 2 kb, the window size was adjusted to $(\text{gene length})/100$, generating 100 windows of TE density values. Finally, the average value for each window position was calculated across all genes, excluding invalid values from the computation.

Phylogenetic tree construction and divergence time estimation

We constructed a phylogenetic tree of 13 wheat subspecies with barley cv. Morex v.3⁹³ as the outgroup using single-copy gene families. Protein sequences of all single-copy genes (corresponding to the longest transcript) were aligned by MAFFT⁹⁴ (v.7.526) and used to construct the phylogenetic trees by IQ-TREE⁹⁵ (v.2.3.5) with `-AUTO`. Phylogenetic trees were constructed using Astral⁹⁶ (v.5.7.8) with `-t 3`.

Divergence time was estimated using MCMCtreeR⁹⁷ (v.4.10.7), referenced in the divergence times of barley versus *T. turgidum* (10.3–13.6 Ma) from fossil records available at Timetree^{598–101} and wild emmer wheat versus cultivated tetraploid wheat (10,400 years ago). The phylogenetic tree of 13 tetraploid wheat and 34 hexaploid wheat accessions was constructed with barley as the outgroup. Sequence data for 34 hexaploid wheat accessions were downloaded from published studies^{24,25,79,102–107}.

NLR genes identification and classification

On the basis of the genome sequences of 13 tetraploid wheat and 34 hexaploid wheat accessions, the nucleotide-binding site leucine-rich repeat (NLR or NBS-LRR) sites were identified by NLR-Annotator¹⁰⁸ (v.2.0). Subsequently, we compared the coordinates of identified NLR sites with the genome-annotation file in GFF3 format and genes with NLR sites were identified as *NLR* genes. In the case of two or more annotated transcripts for a gene, the longest transcript was prioritized for subsequent analysis. Furthermore, protein domains and integrated domains were predicted to divide the *NLR* genes into different classes using InterProScan⁸² (v.5.73; `-dp -iprlookup`). With the classification of *Arabidopsis* as reference²⁷, *NLR* genes were classed into three classes based on the presence of coiled-coil, or resistance to powdery mildew 8 domains: coiled-coil domain-containing CNL-type (*CNL*) genes, *NL* genes (only including at least one NBS and LRR domain) and RESISTANCE TO POWDERY MILDEW 8 (RPW8)-domain-containing RNL-type (*RNL*) genes. Gene families (containing *NLR* genes) were filtered out to form the pangenome of *NLR* genes in tetraploid wheats. Further, compared to hexaploid wheats, the tetraploid gene families with *NLR* genes that could not correspond to the hexaploid genes were defined as specific *NLR* gene families of tetraploids. The distribution of the specific *NLR* gene families of tetraploids and hexaploids was plotted as a heatmap using the pheatmap R package (<https://cran.r-project.org/web/packages/pheatmap/index.html>).

Subgenomic asymmetry analysis

GeneTribe¹⁰⁹ (v.1.2.1) was used to identify syntenic homologous gene pairs between A and B subgenomes based on the longest transcript and protein sequence for each chromosome separately. On the basis of alignment results, GeneTribe¹⁰⁹ (v.1.2.1) classified wheat gene relationships between A and B genes into three types: reciprocal best hit refers to two genes in a syntenic block being each other's best match; singleton denotes genes with no homologous matches; single best hit represents a one-way best match in a syntenic block. Single best hit and singleton genes can reflect the biased genomic distribution of A and B subgenomes among tetraploid wheats and singleton genes showed the specific genes in subgenomes. The R package Rldeogram¹¹⁰ was used to map both asymmetric and symmetric genes on chromosomes. The functions of singleton genes were aligned to *O. sativa* L. proteins (v.IRGSP-1.0, *Oryza sativa* - Ensembl Genomes 61) with BLAST+ (v.2.16.0)⁹² and the best hit (order of prioritization: high identity, high score, low *E*-value) was used to define the function of it.

Graph-based pangenome construction

Using the Kronos genome as reference, the other 11 genomes were aligned using AnchorWave⁶² (v.1.2.6) with the `proali` command. `-R1-Q1`. Alignment results were converted to VCF format using the MAFToG-VCFPlugin of TASSEL¹¹¹ (v.5.2.93) and all VCF files were merged by

BCFtools¹¹² (v.1.20). Then, all variants of VCF files were normalized using the left alignment algorithm implemented in vt¹¹³ (v.0.57721) tool. PanPop¹¹⁴ (v.NC2024) was used to filter the redundant variants and merge the similar multi-allelic sites. From filtered results, we used the vg toolkit¹¹⁵ (v.1.56.0) for graph-based genome construction. IBS analyses were performed using PLINK¹¹⁶ (v.1.9) based on all SVs.

Variant calling and filtering

Clean whole-genome sequencing reads for 621 accessions generated in this study and 115 from published data⁸ were mapped to the Kronos genome using BWA-MEM2 (v.2.3)¹¹⁷ with default parameters. SNPs and small indels (length <50 bp) were genotyped using the LUSH-DNASeq-pipeline¹¹⁸ (v.2.2.0). Only bi-allelic variants were retained and SNPs were filtered using QD < 2.0, SOR > 3.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, ReadPosRankSum < -8.0. Indels were filtered using QD < 2.0, SOR > 10.0, FS > 200.0, MQRankSum < -12.5, ReadPosRankSum < -8.0.

To identify SVs in the 736 accessions, Illumina short reads were mapped to the tetraploid wheat graph-based pangenome using the vg giraffe (v.1.56.0)¹¹⁹ pipeline with default parameters. SVs were genotyped using Paragraph¹²⁰ (v.2.3) with default parameters. VCF files were merged using BCFtools¹¹² (v.1.20) and bi-allelic SVs were retained. SnpEff¹²¹ (v.5.2) was used for annotation and predicting the effects of identified SNPs and small indels. The variants with a missing rate >20% were discarded. Finally, all VCFs were imputed using Beagle¹²² (v.5.4) with default parameters.

Phylogenetic tree and population structure analysis

First, variants were removed by linkage-disequilibrium-based pruning with a window size of 50 SNPs, window step of ten SNPs and r^2 threshold of 0.2 using PLINK¹¹⁶ (v.1.9). Then, we selected 100,000 random variants for phylogeny construction. RAxML¹²³ (v.8.2.12) was used to construct the maximum-likelihood tree with 100 bootstraps in the GTRGAMMA model. Phylogenetic trees were visualized and annotated using Interactive Tree of Life (iTOL) (<https://itol.embl.de>). Ancestry coefficient analysis was conducted using ADMIXTURE¹²⁴ (v.1.3.0), initially with K ranging from 2 to 20, of which $K = 10$ was subsequently chosen. Linkage-disequilibrium decay distance was calculated using PopLDdecay¹²⁵ (v.3.42). Principal component analysis was performed with GCTA¹²⁶ (v.1.94.1).

Genomic selective-sweep analysis

Nucleotide diversity (π), population divergence (F_{st}) and XPCLR scores were used to identify selective sweeps based on high-quality SNPs and small indels. For π and F_{st} , VCFtools¹²⁷ (v.0.1.16) was used with 500-kb sliding windows and 50-kb step size. XPCLR scores were calculated using XPCLR¹²⁸ (v.1.1.2) with the same windows and step size. The R package GenWin¹²⁹ was used to normalize and detect the boundary of selection regions. To identify selected SVs, we compared the allele frequency of each SV between wild emmer wheat and domesticated tetraploid wheat to identify selected SVs during domestication and between domesticated tetraploid wheat and free-threshing tetraploid wheat for selected SVs during improvement. The SV with an allele frequency ≥ 0.5 in at least one population and ≤ 0.5 at least one population simultaneously was used to determine its statistical significance. Fisher's exact test was used to determine significance and the significance threshold for selected SVs was set as $P < 0.05$. If both the P value between wild emmer wheat versus domesticated emmer wheat and domesticated emmer wheat versus free-threshing was < 0.05 , the SV was identified as the selected SV during domestication and improvement continuously. Candidate genes were extracted using BEDTools¹³⁰ (v.2.31.1).

Genome-wide association studies

All high-quality SVs, SNPs and small indels with minor allele frequency (MAF) ≥ 0.05 were retained. All 373 accessions (WEW, DEW and DW)

and 241 accessions (DEW and DW) were used for GWAS analysis using GEMMA¹³¹ (v.0.98.3) software with a mixed linear model, respectively. Kinship matrices were calculated using the command `gemma -gk 2`. The threshold for GWAS was determined by a uniform threshold of $0.05/n$ and $1/n$, where n represents the number of input variants. Candidate quantitative trait loci (QTLs) and genes were identified through a multistep procedure^{132,133}. First, within a 1-Mb region corresponding to the linkage-disequilibrium decay distance, the signal with the lowest P value was designated as the leading SNP, small indel or SV. Next, regions extending 1 Mb upstream and downstream of the leading variant, containing at least two statistically significant loci, were defined as candidate QTLs. Finally, genes located within the linkage-disequilibrium regions corresponding to these QTLs were considered as candidate genes.

Heritability estimation

The proportion of phenotypic variance explained by genetic variants was estimated using LDAK¹³⁴ (v.5.2) with the LDAK-thin model:

$$E[h_j^2] = \tau w_j [f_j(1 - f_j)]^{(1 + \alpha)}$$

where w_j is the weighting for each variant, f_j is its MAF and τ are the corresponding coefficients and are estimated from the variant. All SNPs, indels and SVs were respectively first pruned to exclude nearby variants in perfect linkage disequilibrium using `-window-prune 0.98 -window-kb 100`. The best power parameter α was estimated for each trait by trying multiple values between -0.5 and 1. To filter out negative estimates, we added `-constrain YES`.

Transgenic wheat plants construction

For heterologous expression, complementary DNA prepared from seedling leaves of the tetraploid wheat cultivars Kronos and Zavitan was used as template to amplify the open reading frames of *HAT14-B* and *TtGRAS*. These were then inserted into the pMWB122 vector to generate the *Ubi_{pro}:HAT14* and *Ubi_{pro}:TtGRAS* expression constructs driven by the *Ubiquitin* promoter from maize¹³⁵. For RNAi lines, the 125-bp fragment with the approximately 95% highest *HAT14 A/B/D* sequence similarity was inserted into the pWMB006 vector. The resulting DNA fragments were then inserted into the pCAMBIA3301 vector, generating RNAi constructs. All vectors carrying the desired constructs were introduced into *Agrobacterium tumefaciens* EHA105 and transformed into wheat cv. Fielder via *Agrobacterium*-mediated transformation¹³⁶. Putative transformants were recovered on media containing phosphinothricin at 10 mg l⁻¹, survivors were advanced to soil and transcript levels were quantified by quantitative reverse transcription PCR (qRT-PCR) (primers are listed in Supplementary Table 23). The T_2 generation of transgenic lines was used for phenotyping. Phenotypic evaluation of transgenic lines was conducted (Supplementary Note 1).

RNA isolation and qRT-PCR

For gene-expression assays, total RNA was isolated using the HiPure Plant RNA Mini Kit (R4151; Guangzhou Magen Biotechnology) and 2 μ g of this RNA were reverse transcribed into cDNA using the HiScript II 1st Strand cDNA Synthesis Kit (R4151; Vazyme Biotech). The qRT-PCR was conducted with the ChamQ Universal SYBR qPCR Master Mix (Q711-02; Vazyme Biotech) in accordance with the instructions provided by the manufacturer. Relative expression levels were calculated on the basis of the $2^{-\Delta\Delta Ct}$ method¹³⁷ and were normalized to *Actin1* (*TraesCS1A02G020500*). The experiments incorporated three technical replicates, with each replicate consisting of three plants. All primers used for qRT-PCR are listed in Supplementary Table 23.

Subcellular localization

For the localization assays of *HAT14-B^{Hap1}* and *HAT14-B^{Hap2}*, the pJIT163:GFP expression vector, containing the cauliflower mosaic

virus (CaMV) 35S promoter and harboring the single open reading frame of either *HAT14-B^{Hap1}* or *HAT14-B^{Hap2}*, was introduced into wheat cv. Fielder mesophyll protoplasts through polyethylene glycol-mediated transformation¹³⁸. The empty vector was transformed independently as a control. After incubation for 18 h at 23 °C in darkness, fluorescence was detected by a confocal microscope (FluoView FV3000, Olympus).

For GFP detection, the excitation laser wavelength was 488 nm with emission detected between 500 nm and 540 nm. For chlorophyll autofluorescence, excitation used a 640 nm laser and emission was detected from 650 nm to 750 nm. Laser intensity was controlled by adjusting the acousto-optical tunable filter to 1–4% and visualized with the image-analysis software supplied by the manufacturer.

Dual-luciferase reporter assays

Dual-luciferase reporter assays were carried out¹³⁸. The reporter construct was created by inserting approximately 2-kb promoters of *HAT14-B* and *TtGRAS* into the pGreenII 0800-LUC vector¹³⁹. The effector constructs were generated by cloning the coding sequences of *HAT14-B^{Hap1}* into the pGreenII 62-SK vector¹³⁹, which was driven by the CaMV 35S promoter. *A. tumefaciens* GV3101 was transformed with each respective construct and recombinant strains were then used to co-infiltrate the leaves of 4-week-old *N. benthamiana* plants with the reporter and effector plasmids.

Luciferase levels were measured using a Dual-Luciferase Reporter Assay System (Promega) according to the manufacturer's instructions. The normalized data are presented as the ratio of the luminescent signal intensity of the LUC reporter to that of the internal control REN reporter (*35S_{pro}::REN*) from three independent biological samples.

Electrophoretic mobility shift assay

The coding regions of *HAT14-B^{Hap1}* and *HAT14-B^{Hap2}* were cloned into pMAL-c5X to generate MBP–HAT14-B fusion vectors, respectively. The construct was validated by PCR and transferred into *Escherichia coli* BL21 (DE3) for recombinant protein expression. Biotin-labeled or unlabeled probes containing the AT-containing motif were synthesized by Sangon Biotech and unlabeled probes were used as competitors. Electrophoretic mobility shift assays were performed using the LightShift Chemiluminescent EMSA Kit (Thermo Fisher Scientific) following the manufacturer's protocol. Primers are listed in Supplementary Table 23.

Statistical analyses

Statistical details (for example, tests used, exact *n* values and significance thresholds) are reported in the figure legends. R (v.4.2.1) was applied for analysis unless otherwise indicated and any additional software used is documented in the relevant subsections.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Genome assemblies and HC gene annotations generated in this study have been archived in the Tetraploid Wheat Pan-Genome Database (<http://omicsplant.cn/TetraploidWheatPanGenomeDBs/>) and are available via Zenodo at <https://doi.org/10.5281/zenodo.18466336> (ref. 140). All raw sequencing reads generated from HiFi, ONT, Hi-C, RNA-seq, Iso-seq, ChIP-seq and whole-genome resequencing, as well as assembled genome sequences, have been deposited at the National Genomics Data Center under accession [PRJCA043839](https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA043839) (GSA: [CRA029005](https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA043839)) (<https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA043839>). Seed of 12 tetraploid wheat accessions assembled in this study are available from the corresponding author upon request for research purposes and their germplasm identifiers in gene bank are provided in Supplementary Table 1. Source data are provided with this paper.

Code availability

Code associated with this project is available via GitHub at <https://github.com/Lucio-Yang/Pan-genome> and via figshare at <https://doi.org/10.6084/m9.figshare.32213433> (ref. 141).

References

- Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
- Rautiainen, M. et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* **41**, 1474–1482 (2023).
- Hu, J. et al. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biol.* **25**, 107 (2024).
- Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
- Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, btac808 (2023).
- Alonge, M. et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 258 (2022).
- Song, B. et al. AnchorWave: sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication. *Proc. Natl Acad. Sci. USA* **119**, e2113075119 (2022).
- Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
- Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
- 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).
- Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
- Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
- Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 275 (2019).
- da Cruz, M. H. P., Domingues, D. S., Saito, P. T. M., Paschoal, A. R. & Bugatti, P. H. TERL: classification of transposable elements by convolutional neural networks. *Brief. Bioinform.* **22**, bbaa185 (2020).
- Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
- Shumate, A., Wong, B., Pertea, G. & Pertea, M. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. *PLOS Comput. Biol.* **18**, e1009730 (2022).
- Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
- Li, H. Protein-to-genome alignment with minimap. *Bioinformatics* **39**, btad014 (2023).
- Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, ii215–ii225 (2003).
- Lomsadze, A., Burns, P. D. & Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res.* **42**, e119 (2014).
- Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).

77. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EvidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
78. Feng, J.-W. et al. A haplotype-resolved pangenome of the barley wild relative *Hordeum bulbosum*. *Nature* **645**, 429–438 (2025).
79. Liu, S. et al. A telomere-to-telomere genome assembly coupled with multi-omic data provides insights into the evolution of hexaploid bread wheat. *Nat. Genet.* **57**, 1008–1020 (2025).
80. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
81. UniProt Consortium. UniProt: the universal protein knowledgebase in 2025. *Nucleic Acids Res.* **53**, D609–D617 (2025).
82. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
83. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
84. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
85. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
86. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
87. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
88. Vollger, M. R., Kerpeldjiev, P., Phillippy, A. M. & Eichler, E. E. StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps. *Bioinformatics* **38**, 2049–2051 (2022).
89. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
90. Wickham, H., Navarro, D. & Pedersen, T. L. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).
91. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).
92. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
93. Mascher, M. Pseudomolecules and annotation of the third version of the reference genome sequence assembly of barley cv. Morex [Morex V3]. *e!DAL Plant Genomics and Phenomics Research Data Repository* <https://doi.org/10.5447/ipk/2021/3> (2020).
94. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
95. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
96. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**, 153 (2018).
97. Puttick, M. N. MCMCTreeR: functions to prepare MCMCTree analyses and visualize posterior ages on trees. *Bioinformatics* **35**, 5321–5322 (2019).
98. Kumar, S. et al. TimeTree 5: an expanded resource for species divergence times. *Mol. Biol. Evol.* **39**, msac174 (2022).
99. Miller, A. D. et al. Phylogenetic analyses reveal extensive cryptic speciation and host specialization in an economically important mite taxon. *Mol. Phylogenet. Evol.* **66**, 928–940 (2013).
100. Bernhardt, N., Brassac, J., Kilian, B. & Blattner, F. R. Dated tribe-wide whole chloroplast genome phylogeny indicates recurrent hybridizations within Triticeae. *BMC Evol. Biol.* **17**, 141 (2017).
101. Middleton, C. P. et al. Sequencing of chloroplast genomes from wheat, barley, rye and their relatives provides a detailed insight into the evolution of the Triticeae tribe. *PLoS ONE* **9**, e85761 (2014).
102. Shi, X. et al. Comparative genomic and transcriptomic analyses uncover the molecular basis of high nitrogen-use efficiency in the wheat cultivar Kenong 9204. *Mol. Plant* **15**, 1440–1456 (2022).
103. Jia, J. et al. Genome resources for the elite bread wheat cultivar Aikang 58 and mining of elite homeologous haplotypes for accelerating wheat improvement. *Mol. Plant* **16**, 1893–1910 (2023).
104. Athiyannan, N. et al. Long-read genome sequencing of bread wheat facilitates disease resistance gene cloning. *Nat. Genet.* **54**, 227–231 (2022).
105. Li, G. et al. Genomic analysis of Zhou8425B, a key founder parent, reveals its genetic contributions to elite agronomic traits in wheat breeding. *Plant Commun.* **6**, 101222 (2025).
106. Sato, K. et al. Chromosome-scale genome assembly of the transformation-amenable common wheat cultivar ‘Fielder’. *DNA Res.* **28**, dsab008 (2021).
107. Guo, W. et al. Origin and adaptation to high altitude of Tibetan semi-wild wheat. *Nat. Commun.* **11**, 5085 (2020).
108. Steuernagel, B. et al. The NLR-Annotator tool enables annotation of the intracellular immune receptor repertoire. *Plant Physiol.* **183**, 468–482 (2020).
109. Chen, Y. et al. A collinearity-incorporating homology inference strategy for connecting emerging assemblies in the Triticeae tribe as a pilot practice in the plant pangenomic era. *Mol. Plant* **13**, 1694–1708 (2020).
110. Hao, Z. et al. Rldeogram: drawing SVG graphics to visualize and map genome-wide data on the ideograms. *PeerJ Comput. Sci.* **6**, e251 (2020).
111. Bradbury, P. J. et al. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
112. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
113. Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants. *Bioinformatics* **31**, 2202–2204 (2015).
114. Zheng, Z. et al. A sequence-aware merger of genomic structural variations at population scale. *Nat. Commun.* **15**, 960 (2024).
115. Garrison, E. et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
116. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
117. Vasimuddin, M., Misra, S., Li, H. & Aluru, S. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* 314–324 (IEEE, 2019).
118. Wang, T. et al. Fast and accurate DNaseq variant calling workflow composed of LUSH toolkit. *Hum. Genom.* **18**, 114 (2024).
119. Sirén, J. et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374**, abg8871 (2021).
120. Chen, S. et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol.* **20**, 291 (2019).
121. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80–92 (2012).
122. Browning, B. L., Tian, X., Zhou, Y. & Browning, S. R. Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* **108**, 1880–1890 (2021).
123. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

124. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
125. Zhang, C., Dong, S.-S., Xu, J.-Y., He, W.-M. & Yang, T.-L. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786–1788 (2019).
126. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
127. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
128. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402 (2010).
129. Beissinger, T. M., Rosa, G. J. M., Kaeppler, S. M., Gianola, D. & de Leon, N. Defining window-boundaries for genomic analyses using smoothing spline techniques. *Genet. Sel. Evol.* **47**, 30 (2015).
130. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
131. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
132. Yang, G. et al. Combined GWAS and eGWAS reveals the genetic basis underlying drought tolerance in emmer wheat (*Triticum turgidum* L.). *New Phytol.* **242**, 2115–2131 (2024).
133. Dossa, K. et al. The genetic basis of drought tolerance in the high oil crop *Sesamum indicum*. *Plant Biotechnol. J.* **17**, 1788–1803 (2019).
134. Speed, D., Holmes, J. & Balding, D. J. Evaluating and improving heritability models using summary statistics. *Nat. Genet.* **52**, 458–462 (2020).
135. Pan, W. et al. Efficient gene disruption in polyploid genome by Cas9-Trex2 fusion protein. *J. Integr. Plant Biol.* **67**, 7–10 (2025).
136. Ishida, Y., Hiei, Y. & Komari, T. in *Advances in Wheat Genetics: From Genome to Field* (eds Ogihara, Y. et al.) 167–173 (Springer, 2015).
137. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method. *Methods* **25**, 402–408 (2001).
138. Mao, H. et al. Variation in cis-regulation of a NAC transcription factor contributes to drought tolerance in wheat. *Mol. Plant* **15**, 276–292 (2022).
139. Hellens, R. P. et al. Transient expression vectors for functional genomics, quantification of promoter activity and RNA silencing in plants. *Plant Methods* **1**, 13 (2005).
140. Xu, D. Genome assemblies of 12 tetraploid wheat accessions. *Zenodo* <https://doi.org/10.5281/zenodo.18466336> (2026).
141. Song, B. A pangenome of tetraploid wheat reveals the genetic architecture underlying domestication and genomic diversity for breeding. *figshare* <https://doi.org/10.6084/m9.figshare.32213433> (2026).

Acknowledgements

We thank N. Stein (Leibniz Institute of Plant Genetics and Crop Plant Research), F. Lu (Chinese Academy of Sciences) and H. Sun (Xi'an Jiaotong University) for their valuable suggestions and comments. We also thank the High-Performance Computing Center of Peking University Institute of Advanced Agricultural Science, Shandong Laboratory for Advanced Agricultural Sciences for providing computational resources in this work.

Author contributions

B.S. and X.N. designed the project. B.S., X.N., H.M., X.W.D., Q.S. and S.C. supervised the execution and completion of the project. J.B., G.Y., D.X., Y.Z., Z.Q., B.C., F.L., J.W., C.Y., Y.C., S.L., H.H., S.W. and S.D. performed bioinformatics analyses. Y.J., G.Z., Y.P., Y.L. and Q.Z. managed the fieldwork and prepared the samples. S.Z., Y.W. and M.J. conducted the genetic and molecular experiments. B.S., X.N., H.M., J.B., G.Y., D.X. and Y.Z. wrote the paper, with input from all authors. All authors have read the final paper and approve its submission.

Funding

This project was funded by Key R&D Program of Shandong Province 2024CXPT072 awarded to B.S., the project SYS202206 supported by Shandong Provincial Natural Science Foundation awarded to B.S. and X.W.D., the Major Project on Agricultural Bio-breeding of China no. 2023ZD0402601 awarded to X.N., the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM706) awarded to H.M. and National Natural Science Foundation of China no. 32272103 awarded to Y.L., no. 32372100 awarded to X.N. and no. U22A20457 awarded to X.N.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41588-026-02678-9>.

Correspondence and requests for materials should be addressed to Shisheng Chen, Qing Sang, Xing Wang Deng, Hude Mao, Xiaojun Nie or Baoxing Song.

Peer review information *Nature Genetics* thanks Alison Bentley and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- n/a

Confirmed

☐

☒

The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

☐

☒

A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐

☒

The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☒

☐

A description of all covariates tested

☐

☒

A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐

☒

A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐

☒

For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒

☐

For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒

☐

For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐

☒

Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

No software used for data collection.

Data analysis

Genome assembly: SMRT Link v13.1, dorado v0.4.1, Fastp v0.23.2, NanoFilt v2.8.0, Kmerfreq (no versions, <https://github.com/fanagislab/kmerfreq>), GCE v1.0.2, Genomescope v1.0, findGSE v0.1.0, hifiasm v0.22.0-r689, minimap2 v2.24, YaHS v1.2a.1, Juicebox v1.11.08, HiC-Pro v3.1.0, BUSCO v5.4.6, Merqury v1.3, LTR_retriever v2.9.0, Verkko v2.0, NextDenovo v2.5.2, RagTag v2.1.0, AnchorWave v1.2.6, NextPolish v1.4.1, Bowtie2 v2.5.1, SAMtools v1.10, MACS3 v3.0.1, Deeptools v3.5.5, StainedGlass v0.6, NanoPack v1.1.1.
Genome annotation: Tandem Repeats Finder v4.09, EDTA v2.0.1, RepeatMasker v4.1.1, STAR v2.7.10b, StringTie v3.0.0, pbmm2 v1.17.0, GMAP v2024-11-20, TransDecoder v5.7.1, cufflinks v2.2.1, miniprot v0.12, Augustus v3.5.0, GeneMark-ET, v4.72_lic, EvidenceModeler v2.1.0, PASA v2.5.3, DIAMOND v2.1.8, InterProScan v5.73.
Pangenome analysis: DAGchainer (no version, https://vcru.wisc.edu/simonlab/bioinformatics/programs/dagchainer/dagchainer_documentation.html), SynPan (no versions, <https://github.com/dongyawu/PangenomeEvolution?tab=readme-ov-file>), BLAST v2.16.0, MAFFT v7.526, IQ-TREE2 v2.3.5, Astral v5.7.8, MCMCTreeR v4.10.7, NLR-Annotator v2.0, GeneTribe v1.2.1, TASSEL v5.2.93, PLINK v1.9, PanPop vNC2024, vt v0.57721, AnchorWave v1.2.6, BWA v0.7.17, OrthoFinder v2.5.5, CAFE v5, bcftools v1.20.
Population analysis: BWA-MEM2 v2.3, LUSH_toolkit-HC v2.2.0, vg v1.56.0, Paragraph v2.3, SnpEff v5.2, Beagle v5.4, RAXML v8.2.12, ADMIXTURE v1.3.0, PopLDdecay v3.42, GCTA v1.94.1, Cytoscape v3.6, CytoCluster plugin v2.1.0, AutoAnnotate plugin v1.5.2, XPCLR v1.1.2, VCftools v0.1.16, BEDTools v2.31.1, GenWin 1.0, GEMMA v0.98.3, LDAK v5.2, ABBABABAwindows.py (no versions, https://github.com/simonhmartin/genomics_general).
Code associated with this project is available at Github (<https://github.com/Lucio-Yang/Pan-genome>) and Figshare (<https://doi.org/10.6084/m9.figshare.32213433>).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Genome assemblies and high-confidence gene annotations generated in this study have been archived in the Tetraploid Wheat Pan-Genome Database (<http://omicsplant.cn/TetraploidWheatPangenomeDBs/>) and Zenodo (<https://doi.org/10.5281/zenodo.18466336>).

All raw sequencing reads generated from HiFi, ONT, Hi-C, RNA-seq, Iso-seq, ChIP-seq and whole-genome resequencing, as well as assembled genome sequences, have been deposited at the National Genomics Data Center (NGDC) under accession PRJCA043839 (<https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA043839>).

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

Reporting on race, ethnicity, or other socially relevant groupings

Population characteristics

Recruitment

Ethics oversight

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Data exclusions

Replication

Randomization

Blinding

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☐	☒ Plants

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	The peptide of the specific N-terminal tail of TaCENH3, 'ARTKHPAVRKTk', was expressed in the pET28b vector fused to a His-tag in Escherichia coli BL-21. An anti-rabbit polyclonal antibody against this peptide was prepared and purified using rabbit serum by AtaGenix (Wuhan, China).
Validation	The TaCENH3 antibody has been widely used for ChIP on wheat in previous publications (Walkowiak et al., 2020; Liu et al., 2023).

Dual use research of concern

Policy information about [dual use research of concern](#)

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes
☒	☐ Public health
☒	☐ National security
☒	☐ Crops and/or livestock
☒	☐ Ecosystems
☒	☐ Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes
☒	☐ Demonstrate how to render a vaccine ineffective
☒	☐ Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	☐ Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	☐ Increase transmissibility of a pathogen
☒	☐ Alter the host range of a pathogen
☒	☐ Enable evasion of diagnostic/detection modalities
☒	☐ Enable the weaponization of a biological agent or toxin
☒	☐ Any other potentially harmful combination of experiments and agents

Plants

Seed stocks	All tetraploid wheat accessions were kept in Peking University Institute of Advanced Agricultural Sciences.
Novel plant genotypes	Not applicable.
Authentication	Not applicable.

ChIP-seq

Data deposition

- ☒ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- ☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links

May remain private before publication.

ChIP-seq data have been deposited in the Genome Sequence Archive (GSA) database (<https://ngdc.cncb.ac.cn/gsa/>) under accession PRJCA043839.

Files in database submission

IG77365_CENH3_1_paired.fq.gz
 IG77365_CENH3_2_paired.fq.gz
 IG77365_Input_1_paired.fq.gz
 IG77365_Input_2_paired.fq.gz
 IG99236_CENH3_1_paired.fq.gz
 IG99236_CENH3_2_paired.fq.gz
 IG99236_Input_1_paired.fq.gz
 IG99236_Input_2_paired.fq.gz
 Kronos_CENH3_1_paired.fq.gz
 Kronos_CENH3_2_paired.fq.gz
 Kronos_Input_1_paired.fq.gz
 Kronos_Input_2_paired.fq.gz
 NU00021_CENH3_1_paired.fq.gz
 NU00021_CENH3_2_paired.fq.gz
 NU00021_Input_1_paired.fq.gz
 NU00021_Input_2_paired.fq.gz
 NU01905_CENH3_1_paired.fq.gz
 NU01905_CENH3_2_paired.fq.gz
 NU01905_Input_1_paired.fq.gz
 NU01905_Input_2_paired.fq.gz
 NU01954_CENH3_1_paired.fq.gz
 NU01954_CENH3_2_paired.fq.gz
 NU01954_Input_1_paired.fq.gz
 NU01954_Input_2_paired.fq.gz
 PI294478_CENH3_1_paired.fq.gz
 PI294478_CENH3_2_paired.fq.gz
 PI294478_Input_1_paired.fq.gz
 PI294478_Input_2_paired.fq.gz
 PI330553_CENH3_1_paired.fq.gz
 PI330553_CENH3_2_paired.fq.gz
 PI330553_Input_1_paired.fq.gz
 PI330553_Input_2_paired.fq.gz
 XM000719_CENH3_1_paired.fq.gz
 XM000719_CENH3_2_paired.fq.gz
 XM000719_Input_1_paired.fq.gz
 XM000719_Input_2_paired.fq.gz
 XM000834_CENH3_1_paired.fq.gz
 XM000834_CENH3_2_paired.fq.gz
 XM000834_Input_1_paired.fq.gz
 XM000834_Input_2_paired.fq.gz
 XM001097_CENH3_1_paired.fq.gz
 XM001097_CENH3_2_paired.fq.gz
 XM001097_Input_1_paired.fq.gz
 XM001097_Input_2_paired.fq.gz
 Zavitan_CENH3_1_paired.fq.gz
 Zavitan_CENH3_2_paired.fq.gz

Genome browser session
(e.g. [UCSC](#))

Zavitan_Input_1_paired.fq.gz
Zavitan_Input_2_paired.fq.gz

N/A

Methodology

Replicates

N/A. All samples were obtained from 3-week-old seedlings.

Sequencing depth

All files were 150 bp pair-end sequencing.
IG77365_CENH3: 64325652 bp
IG77365_Input: 39334628 bp
IG99236_CENH3: 59123698 bp
IG99236_Input: 38679324 bp
Kronos_CENH3: 92965618 bp
Kronos_Input: 24052518 bp
NU00021_CENH3: 59731688 bp
NU00021_Input: 43652906 bp
NU01905_CENH3: 67529096 bp
NU01905_Input: 60880106 bp
NU01954_CENH3: 78116698 bp
NU01954_Input: 25576872 bp
PI294478_CENH3: 71835060 bp
PI294478_Input: 35818206 bp
PI330553_CENH3: 64361608 bp
PI330553_Input: 42423566 bp
XM000719_CENH3: 71307552 bp
XM000719_Input: 36633660 bp
XM000834_CENH3: 57130194 bp
XM000834_Input: 46030068 bp
XM001097_CENH3: 72485872 bp
XM001097_Input: 35452426 bp
Zavitan_CENH3: 74034868 bp
Zavitan_Input: 23602426 bp

Antibodies

Wheat CenH3 antibody.

Peak calling parameters

MACS3 was used to call CENH3 ChIP-seq peaks using input as a control with the settings: -f BAMPE. Adjacent significant CENH3-enriched regions within 1,000 kb were merged, and the most significant regions were determined as potential centromere regions.

Data quality

Adapters and low-quality reads in ChIP-seq data for 12 accessions were removed using Trimmomatic v0.39 with the parameters "ILLUMINACLIP: adapter.fa:2:30:10 MINLEN:36 SLIDINGWINDOW: 4:20". Quality-controlled reads were aligned to each genome using Bowtie283 with the parameters (--very-sensitive --no-mixed --no-discordant -k 10 -X 1000) and were further filtered by SAMtools v1.10 with the parameters (-F 1804 -f 2 -q 30).

Software

Bowtie2 v2.5.1, SAMtools v1.10, MACS3 v3.0.1, Deeptools v3.5.5, StainedGlass v0.6, BEDtools v2.31.1