Article

Pan-genome bridges wheat structural variations with habitat and breeding

https://doi.org/10.1038/s41586-024-08277-0

Received: 16 January 2024

Accepted: 23 October 2024

Published online: 27 November 2024

Check for updates

Chengzhi Jiao[1,2,9], Xiaoming Xie[3,9], Chenyang Hao[1,9], Liyang Chen[4,9], Yuxin Xie[1], Vanika Garg[5], Li Zhao[1], Zihao Wang[3], Yuqi Zhang[3], Tian Li[1], Junjie Fu[1], Annapurna Chitikineni[5], Jian Hou[1], Hongxia Liu[1], Girish Dwivedi[6,7], Xu Liu[1], Jizeng Jia[1], Long Mao[1], Xiue Wang[2], Rudi Appels[5,8], Rajeev K. Varshney[5]✉, Weilong Guo[3]✉ & Xueyong Zhang[1]✉

Wheat is the second largest food crop with a very good breeding system and pedigree record in China. Investigating the genomic footprints of wheat cultivars will unveil potential avenues for future breeding efforts[1,2]. Here we report chromosome-level genome assemblies of 17 wheat cultivars that chronicle the breeding history of China. Comparative genomic analysis uncovered a wealth of structural rearrangements, identifying 249,976 structural variations with 49.03% (122,567) longer than 5 kb. Cultivars developed in 1980s displayed significant accumulations of structural variations, a pattern linked to the extensive incorporation of European and American varieties into breeding programmes of that era. We further proved that structural variations in the centromere-proximal regions are associated with a reduction of crossover events. We showed that common wheat evolved from spring to winter types via mutations and duplications of the *VRN-A1* gene as an adaptation strategy to a changing environment. We confirmed shifts in wheat cultivars linked to dietary preferences, migration and cultural integration in Northwest China. We identified large presence or absence variations of *pSc200* tandem repeats on the 1RS terminal, suggesting its own rapid evolution in the wheat genome. The high-quality genome assemblies of 17 representatives developed and their good complementarity to the 10+ pan-genomes offer a robust platform for future genomics-assisted breeding in wheat.

Common wheat (AABBDD, *Triticum aestivum* L.), one of the most productive crops, originated through natural hybridization between free-threshing tetraploid wheat (AABB) and *Aegilops tauschii* (DD) in the Middle East about 7,000–8,000 years ago[3–6]. Common wheat spread to Europe and East Asia before being introduced to America, South America and Australia by European colonists, demonstrating the substantial impact of evolutionary limitations in shaping the adaptive environment of bread wheat[7–9]. Wheat expanded from the limited core area of the Fertile Crescent to a vast spectrum of habitats worldwide, becoming the most widely cultivated crop. Introduced to China approximately 3,500–4,000 years ago, wheat prompted a substantial shift in agricultural practices, replacing millet and eventually dominating the agricultural landscape of North China and expanding along the Yangzi River to become a primary food crop after rice[10]. Since 1950, about 3,500 new wheat cultivars have been bred and released in China, among which a subset of cultivars stood out in production and breeding, thus serving as founder genotypes in the national wheat breeding system[11,12]. A platform displaying the genomes and genetic variations among these founders and their new derivatives will help to identify valuable genetic resources, benefiting future plant breeding efforts[13–15].

Complex ecosystems in China, including distinctions such as spring versus winter cultivation, rain-fed versus irrigation farming, and one-crop versus two-crop seasons, have continually driven the creation and retention of rich genetic diversity[11]. The investigation into how wheat varieties have adapted to the vast terrain in China, and their evolution in response to shifting climatic conditions and domestication needs, has drawn the attention of botanists and breeders. The understanding derived from this research holds the potential to unlock key insights into the adaptive capabilities of this pivotal crop[16].

Sociocultural food practices profoundly influence the choices of resources and food production systems[17]. Western preferences tend towards baked goods such as bread, cookies and doughnuts, whereas Asian nations, including China, predominantly consume wheat in the form of steamed bread, stuffed buns, noodles and dumplings. The relationship between food culture and wheat cultivar evolution under human selection, shaped by dramatic shifts in lifestyle, dietary

[1]State Key Laboratory of Crop Gene Resources and Breeding/Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China. [2]National Key Laboratory of Crop Genetics and Germplasm Enhancement and Utilization, Nanjing Agricultural University, Nanjing, China. [3]Frontiers Science Center for Molecular Design Breeding, Key Laboratory of Crop Heterosis and Utilization, Beijing Key Laboratory of Crop Genetic Improvement, China Agricultural University, Beijing, China. [4]Smartgenomics Technology Institute, Tianjin, China. [5]Centre for Crop and Food Innovation, WA State Agricultural Biotechnology Centre, Food Futures Institute, Murdoch University, Murdoch, Western Australia, Australia. [6]Harry Perkins Institute of Medical Research, the University of Western Australia, Murdoch, Western Australia, Australia. [7]Department of Cardiology, Fiona Stanley Hospital, Murdoch, Western Australia, Australia. [8]AgriBio, Centre for AgriBioscience, Department of Economic Development, Jobs, Transport, and Resources, La Trobe University, Bundoora, Victoria, Australia. [9]These authors contributed equally: Chengzhi Jiao, Xiaoming Xie, Chenyang Hao, Liyang Chen. ✉e-mail: rajeev.varshney@murdoch.edu.au; guoweilong@cau.edu.cn; zhangxueyong@caas.cn

alterations and food production, defines genetic principles driving adaptive success that are vital for securing productive and stable future wheat yields.

High-quality genome assembly is crucial for the isolation of valuable genes[18,19]. The establishment of a gold-standard reference genome of Chinese Spring in 2018 has substantially advanced our understanding of the diversity and evolution of the wheat genome through the resequencing of wheat and its progenitors[1,16]. Furthermore, pan-genomes provide a new efficient strategy to address important questions such as heterosis[20] and the uncovering of hidden genomic variations in breeding[21,22].

In this study, 17 common cultivars were sequenced by PacBio in combination with Hi-C technology. These 17 cultivars, including one landraces, 15 modern cultivars and one introduced European cultivar, were selected based on their substantial contribution to modern Chinese wheat breeding, production and their possession of unique traits, in line with the national breeding history of the past 70 years. We also integrated the available wheat genomic resources, such as the 10+ pan-genomes, and the genomes of Fielder, Kariega and Attraktion along with resequencing data[2,11].

At the pan-genome level, we identified a significant number of structural variations (SVs), including presence or absence variations (PAVs), translocations and inversions. We unveiled the integration of introduced European germplasm into modern Chinese wheat breeding by comparing PAV frequencies across cultivars from different decades. In addition, we observed an association between changes in gene copy number and allele types in wheat spread and breeding (for example, the *VRN-A1* gene), indicating that its genome is still rapidly evolving, which was also evidenced by the ongoing deletion of *NOR* copies in breeding[23]. Our analysis also highlighted differences in wheat variety selection between Western and Eastern nations, rooted in diverse dietary preferences and consumption patterns. Furthermore, we detected alien fragments in wheat using the newly assembled genomes of CM42, derived from a cross of CIMMYT synthetic with local modern cultivars[24]. A large PAV block (about 20 Mb) showing a significant reduction of *pSc200* was detected on the de novo-assembled 1RS in the cultivars with 1RS·1BL translocations, which has been widely used in production. Overall, we have presented a wheat pan-genome constructed from newly assembled representative Chinese cultivars and revealed that the dynamic utilization of SVs coevolved with the habitat and food culture of wheat.

Genomes of 17 wheat cultivars

To capture the genomic diversity of common wheat cultivars representing the breeding history spanning 70 years in China, 17 cultivars were selected for de novo genome assembly (Fig. 1a, Extended Data Fig. 1 and Supplementary Table 1), covering lines of different breeding stages: 1950–1960s, 1980–1990s and post-2000s[2,11,25,26] (Fig. 1b and Extended Data Fig. 2). These cultivars complement the genetic diversity from eastern Asia, mostly uncovered by previously available genome assemblies (Fig. 1a). The cultivars were sequenced using PacBio HiFi with an average depth of 30.37× (Supplementary Table 2) and were de novo-assembled to obtain final assembly sizes from 14.6 Gb (NC4) to 15.1 Gb (HD6172), with an average of 14.86 Gb. All 17 cultivars were anchored to the chromosome level using Hi-C with an average depth of 63.82× (Supplementary Table 3 and Supplementary Fig. 1). The average contig N50 of the 17 assemblies was 27.36 Mb, and the average number of gaps was 2,288. An average of 97.38% of the contigs was anchored to the chromosomes. The BUSCO completeness of these assemblies was estimated to be over 98.90% (Supplementary Table 6). Furthermore, the LTR assembly index (LAI) indicated that all of the assemblies reached the 'reference' level (LAI > 10; Extended Data Fig. 2 and Supplementary Table 7). Using a combination of homology-based, ab initio and transcriptome-evidence-based prediction, we identified an average of approximately 153,077 protein-coding genes for each cultivar (Supplementary Table 10). On the basis of various assembly metrics, the 17 assemblies developed were comparable with the four previously published assemblies (Chinese Spring[1], Fielder[27], Kariega[28] and Attraktion[29]) in terms of assembly contiguity, suggesting the high quality of the newly developed assemblies (Extended Data Table 1).

Pan-genome and dynamics of NLR gene families

Next, we performed pan-genome analysis using orthologue identification by integrating the 17 newly developed assemblies along with four previously published assemblies (Chinese Spring[1], Fielder[27], Kariega[28] and Attraktion[29]). On the basis of orthologue analysis, all genes from the 21 genomes were grouped into 170,517 potential gene families, of which 111,955 families (65.66%) comprised members from all 21 cultivars and were thus defined as core gene families. Moreover, 12,486 gene families contained members from 19 to 20 cultivars and were defined as softcore gene families (7.32%). The remaining 46,076 families present in less than 19 cultivars were defined as dispensable genes (27.02%) (Extended Data Fig. 3a,b, Supplementary Table 13 and Supplementary Fig. 3). The saturation curve indicated that the number of total gene families at the whole-genome level approached saturation when $n = 10$ (Supplementary Fig. 3a), suggesting that the current wheat pan-genome is sufficiently representative to encompass all gene families. Furthermore, KEGG enrichment analysis revealed that core genes are predominantly associated with basic metabolism, whereas dispensable genes are mainly involved in pathways such as plant resistance regulation (Supplementary Tables 14 and 15 and Supplementary Fig. 4).

Among R proteins, the nucleotide-binding leucine-rich repeat (NLR) genes form a major protein family. We investigated the copy number changes of NLR genes across genomes from different periods (1950–1960s, 1980–1990s and post-2000s; Supplementary Table 16). Of note, cultivars released in the 1980–1990s, such as XY6 and YM158, exhibited the highest number of NLR genes (Supplementary Fig. 5a). A slight decrease was observed in the copy number of CC-NBARC-LRR (coiled-coil, nucleotide-binding site and leucine-rich repeat) genes over time, whereas changes in other genes, including NBARC-LRR (nucleotide-binding site and leucine-rich repeat), were not substantial (Supplementary Fig. 5b). Furthermore, at the subgenome level, the gene families across A, B and D subgenomes in the pan-genome reached saturation, whereas the NLR gene set remained unsaturated (Supplementary Fig. 5c,d), indicating the highly dynamic PAV of NLR genes due to rapid evolution. This highlights an unsaturated representation of the NLR genes in the pan-genome, indicating the potential to discover novel disease-resistant genes in future wheat-breeding initiatives.

Genome SV landscape

To identify SVs (including PAVs, translocations and inversions), the 17 newly assembled genomes were compared with the Chinese Spring reference. We identified a total of 249,976 SVs, including 119,331 presences and 116,046 absences, 13,550 translocation events and 1,049 inversion events, and 49.03% (122,567) of the SVs were longer than 5 kb and the average gene length was 2.99 kb from the start codon to the stop codon in current assembles (Extended Data Fig. 4 and Supplementary Table 19). Among wheat cultivars released at different periods, the number of SVs increased from older cultivars to the more recent cultivars, with a pronounced accumulation of SVs in cultivars released since the 1980s (Fig. 1c and Supplementary Fig. 6b). Several large inter-chromosomal translocations were identified between cultivars (Fig. 1d, Supplementary Table 20 and Supplementary Fig. 7). The distribution of PAVs on each chromosome showed the highest accumulation level in the B subgenome, followed by the A subgenome and the lowest level in the D subgenome (Fig. 1e, Extended Data Fig. 5 and Supplementary Fig. 6a). In the B subgenome, several chromosomal segments depicted a high density of PAVs. Furthermore, the PAVs appeared

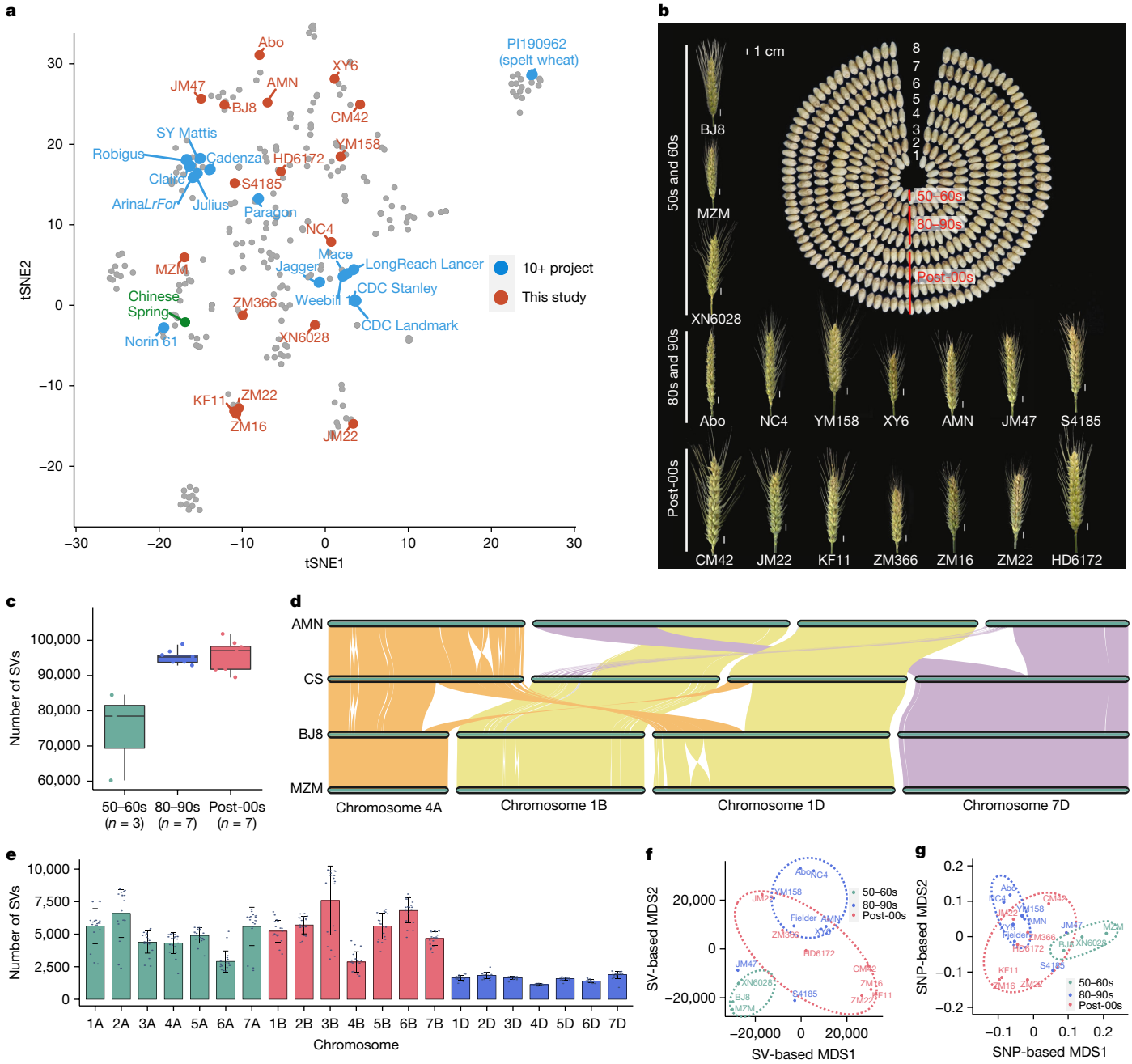


Fig. 1 | Heads and seeds of the 17 wheat cultivars, their assembled genomes and representativeness for local and global wheat diversity. a, *t*-distributed stochastic neighbour embedding (tSNE) analysis of the SNPs from whole-genome shotgun resequencing of 285 worldwide representative germplasms from three previous studies[11,25,26] (grey), 15 lines from the 10+ genome sequencing project[2] (blue), Chinese Spring (green) and our new assemblies (red). **b**, Phenotypes of heads and seeds of the 17 cultivars. **c**, Boxplot of the number of SVs among cultivars released in three stages. **d**, Large structural translocations in wheat cultivars. A translocation is observed between chromosome 1B and chromosome 7D between AMN and Chinese Spring (CS), whereas BJ8 and MZM share a common translocation region between chromosome 4A and chromosome 1D relative to Chinese Spring. **e**, The number of PAVs in the 21 chromosomes. **f**,**g**, Genetic relationships among de novo assembled wheat cultivars revealed by multidimensional scaling (MDS) based on SVs (**f**) and SNPs (**g**).

to be enriched in the promoter regions (Extended Data Fig. 4b), and the number of identified PAVs decreased along with the PAV length (Extended Data Fig. 4c,f).

When tracking changes in SVs over the past 70 years, we observed that a large proportion of SVs showed significant differences among varieties from different stages, with 32.95% of SVs getting nearly fixed in the 1980s due to extensive selection in breeding (Extended Data Fig. 3d,e). Multidimensional scaling analysis revealed a clear separation of wheat varieties released in the 1950s and 1980s, whereas those generated after the 2000s spread along the first dimension and integrated with the other two groups, suggesting the reuse and integration of European diversity in modern Chinese wheat breeding (Fig. 1f,g and Extended Data Fig. 3d,e).

Pericentromeric SVs reduce recombination

As the 17 cultivar assemblies demonstrated good representation of cultivars in China released since 1940s (Fig. 1a,b and Extended Data Fig. 1),

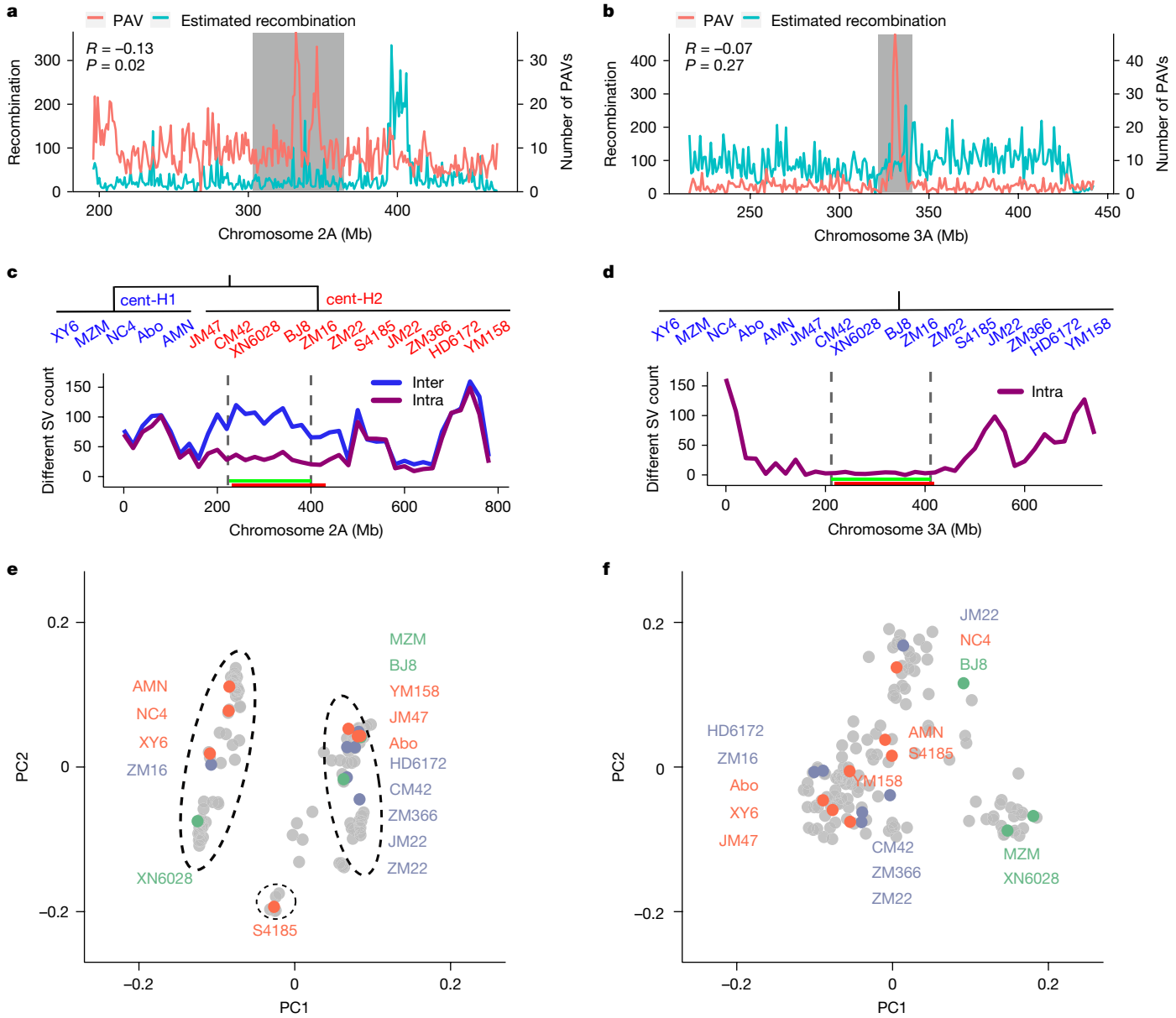


Fig. 2 | PAVs in the centromere-proximal region halted crossover recombination. a,**b**, Correlation of the CRN and PAV number within 100 Mb upstream and downstream of the centromeres (grey shaded region) on chromosome 2A (**a**) and chromosome 3A (**b**). **c**, Two centAHGs of chromosome 2A were grouped among the 17 de novo assemblies, labelled as cent-H1 and cent-H2 (top). The number of different SVs between assembly pairs for the intra-centAHG (purple) and inter-centAHG (blue) groups on chromosome 2A is also shown (bottom). The red bar indicates the centromere-proximal region from 100 Mb upstream to 100 Mb downstream of the centromere. The green bar denotes the centAHG block previously identified[32]. **d**, Only one centAHG was detected on chromosome 3A among the 17 de novo assemblies. **e**,**f**, Principle component analysis (PCA) of resequenced 145 landmarker cultivars based on SVs within 100 Mb upstream and downstream of the centromeres on chromosome 2A (**e**) and chromosome 3A (**f**). The green scatter dots represent the cultivars released in the 1950–1960s; the orange scatter dots represent the cultivars released in 1980–1990s; the blue scatter dots represent the cultivars released post-2000s; and the grey scatter dots represent others.

we used the crossover recombination number (CRN) in the 145 resequenced cultivars to reveal the effect of PAVs on recombination at regions proximal to the centromeres. A high number of PAVs were detected in almost all chromosomes at their core centromere regions (shaded in grey in Fig. 2a,b) compared with the Chinese Spring reference. High accumulation of SVs at regions proximal to centromeres, similar to findings in humans, was significantly associated with reduced recombination and the formation of extensive haplotype blocks across centromeres in wheat[16,30–32] (Fig. 2a,b and Extended Data Fig. 6).

To investigate the impact of SVs associated with the centromeric ancestral haplotype group (centAHG) blocks, we grouped the assemblies by previously characterized centAHGs[32] and profiled the SV frequencies (Fig. 2e,f, Supplementary Table 21 and Supplementary Fig. 7). Results showed significantly higher SV frequencies at the centromere-proximal region for inter-centAHG pairs than for intra-centAHG pairs (Fig. 2c and Supplementary Table 22), whereas no significant difference was observed in the boundary regions of chromosomes (Fig. 2d). Thus, the SVs revealed by the panel of genome assemblies explain the low recombination rates at the centromere-proximal regions with different ancestral haplotypes that are derived from distinct wild emmer lineages. This was also strongly supported by the obvious difference between the D subgenome and the A and B subgenomes (Extended Data Fig. 7).

CRN in breeding history was also strongly affected by differentiation at regions proximal to centromeres. For example, chromosomes 3A and 5A in the A subgenome, with less differentiation, have higher CRN. This is also applicable to explain the extremely low CRN at regions proximal to centromeres on chromosomes 3B and 6B. All of the seven chromosomes in the D subgenome still underwent differentiation at regions proximal to their centromeres, with rapid differentiation on chromosome 7D (Extended Data Fig. 6).

The most notable exception occurred on chromosome 4A, where low PAVs correlated with an extremely low CRN in breeding history. One of the most likely reasons is a 300-Mb introgression across the centromere from *Triticum dicoccoides*, which strongly suppressed chromosome recombination, leading to significant differentiation in the bred cultivars (Extended Data Fig. 6). The exception of chromosome 7A proximal to the centromere can also be explained by this reason[16,33].

Evolution of *VRN-A1* duplications

Next, we focused on the identification of SVs selected during breeding. The proportions of selected SVs from the 1950s to the 1980s and from the 1980s to the 2000s were 13.52% and 13.16%, respectively (Supplementary Fig. 12a). Using both single-nucleotide polymorphisms (SNPs) and SVs to screen for genome-wide loci under selection during wheat breeding, we identified a vernalization gene (*VRN-A1*) that controls the ecotype (spring or winter), and two puroindoline (PIN) genes (*Pina* and *Pinb*) that regulate wheat hardness, presented within regions of extreme F_{ST} values (Supplementary Fig. 12b). Genome-wide association study (GWAS) analysis, using ecotype as the trait, also showed a clear signal at the region containing *VRN-A1* (Fig. 3a).

On the basis of the two SNPs in the coding sequence (CDS) region of the *VRN-A1* gene in the assembled bread wheat cultivars, tetraploid wheat genomes and the barley genome, the gene haplotypes were divided into three subclades, including one spring-type clade and two winter-type clades (Fig. 3b). When analysing the change of the *VRN-A1* allele type in wheat breeding, we observed the appearance of the spring allele in early varieties such as Chinese Spring, Abo and NC4, and the emergence of the two winter alleles in more modern varieties, such as JM22 and HD6172. In addition, genome-wide comparison identified multiple duplications of the region harbouring *VRN-A1*. The copy numbers of *VRN-A1* were associated with the ecotype: the strong winter type, such as JM22, contained higher copies of *VRN-A1* genes than the winter wheat, such as YM158, and the spring wheat, such as NC4 (Fig. 3c, Supplementary Table 23 and Supplementary Fig. 14a–c).

Whole-genome resequencing data also supported the presence of two duplicated regions containing the *VRN-A1* gene in YM158 (Fig. 3d). Thus, we used the resequencing data to estimate the *VRN-A1* copy numbers in a population consisting of 423 hexaploid wheat accessions. The distribution of copy numbers showed that most of these varieties contain one to three copies, whereas a few varieties have more (four or even five; Fig. 3e and Supplementary Fig. 15a). Duplications of the *VRN-A1* gene were also observed during domestication from the tetraploid emmer (wild emmer wheat) to the domesticated emmer wheat and durum wheat, and then to hexaploid wheat (Fig. 3f and Supplementary Table 24).

Along with the spread of wheat from the Middle East to China, the copy number of *VRN-A1* increased, and collections in northern regions carried more duplications than those in southwestern regions in the landraces (Extended Data Fig. 8a). We also noted that the copy number of *VRN-A1* gradually decreased in modern cultivars released in northern China during the past 70 years. The triple duplications were mainly distributed in landraces, whereas the modern cultivars contained mostly two copies with a small proportion of three-copy varieties, suggesting the potential role of *VRN-A1* in new cultivars adapting to the warming climate (Extended Data Fig. 8b and Supplementary Table 25). The warming climate was confirmed by the mean temperature in January in the past 60 years in Henan, the province with the largest wheat production in China (Extended Data Fig. 8c).

After 3 weeks of cold treatment (5–8 °C), the spring cultivars usually showed higher expression at *VRN-A1* (Fig. 3g). However, after full vernalization in the field (from 13 November 2023 to 2 March 2024 in Beijing), we found that the total transcription level of *VRN-A1* genes in winter cultivars was much higher than in spring cultivars (Fig. 3g). This explained why winter cultivars headed earlier than the spring cultivars when they were planted in winter wheat zone II (Huang-huai zone; accounting for 50% and 60% of acreage and production, respectively, in China). Generally, the cultivars with haplotype II and haplotype III (such as JM47 and ZM16) had higher transcription levels than those with haplotype I (such as Chinese Spring, Abo and NC4). Moreover, the II + III duplications usually had higher transcription levels than I + III duplications, such as in XY6, AMN and YM158 with higher expression than CM42. Of note, most of the winter and strong winter cultivars favoured the duplication of Hap-II + III (Fig. 3g). These results indicate that co-expression among these duplicates is much more complex than previously expected, taking into account earlier findings that epigenetic modifications were detected at the *VRN-A1* loci and that SV typically affects gene expression[34,35].

PIN variations and grain hardness

GWAS analysis of grain hardness showed a clear signal in the region containing *Pina* and *Pinb* (Fig. 4a). Two alleles for *Pina* (*Pina-D1a* as the wild-type allele and *Pina-D1b* as the allele causing gene deletion) and three alleles for *Pinb* (*Pinb-D1a* as the wild-type allele, *Pinb-D1b* carrying a non-synonymous mutation and *Pinb-D1u* as the allele carrying a premature stop codon) were identified in the existing assemblies (Fig. 4b,c). Varieties with the wild-type alleles produce softer grains, whereas the mutant alleles produce significantly harder grains (Fig. 4d,e and Supplementary Tables 26 and 27).

When comparing the haplotypes for *Pina* and *Pinb* between ten Western cultivars in the 10+ pan-genome and 17 Chinese cultivars, we observed a higher allelic diversity in Chinese cultivars (Fig. 4 and Supplementary Table 27). For the *Pina* gene, the allele causing gene deletion, *Pina-D1b*, predominantly spreads to the East; and for the *Pinb* gene, the *Pinb-D1b* allele had a large proportion in Europe, whereas the *Pinb-D1u* allele appears in the East, indicating different selection and utilization of PIN alleles between Western and Eastern countries (this might also suggest potential roles of PIN genes in traits other than grain hardness; Fig. 4d,e). Beyond the temperature effect, cooking styles may also influence the geographical distribution of the PIN genes in those regions. The spread of wild-type haplotypes (*Pina-D1a* and *Pinb-D1a*; Fig. 4d,e) in the southern parts of China suggests the selection of soft grains (*Pina-D1a*) to cook boiled and steamed foods (buns and noodles; Extended Data Fig. 9a–d). By contrast, the selection of the mutant haplotype (*Pina-D1b*) reflects that the northern part of China, with the settlement of minority ethnic groups after immigration, kept their unique cooking style for baking and cooking just like those of Europe and the Middle East, opting for cultivars producing hard grains.

Several wheat varieties from Western countries carry an allele with a *Pina* deletion, found only in a single Chinese cultivar: NC4 (Fig. 4c). NC4 was bred and released in the 1980s in Ningxia but remains widely cultivated in northwestern China, which is home to many minority ethnic groups. These communities share a similar food preparation style with the Western and Middle Eastern cultures, favouring baked wheat products, which could explain their preference for this hard wheat variety where the *Pina* gene was fully deleted. By contrast, most of the Han population in China favours steamed and boiled foods, such as steamed bread and noodles, leading to a wider distribution of softer

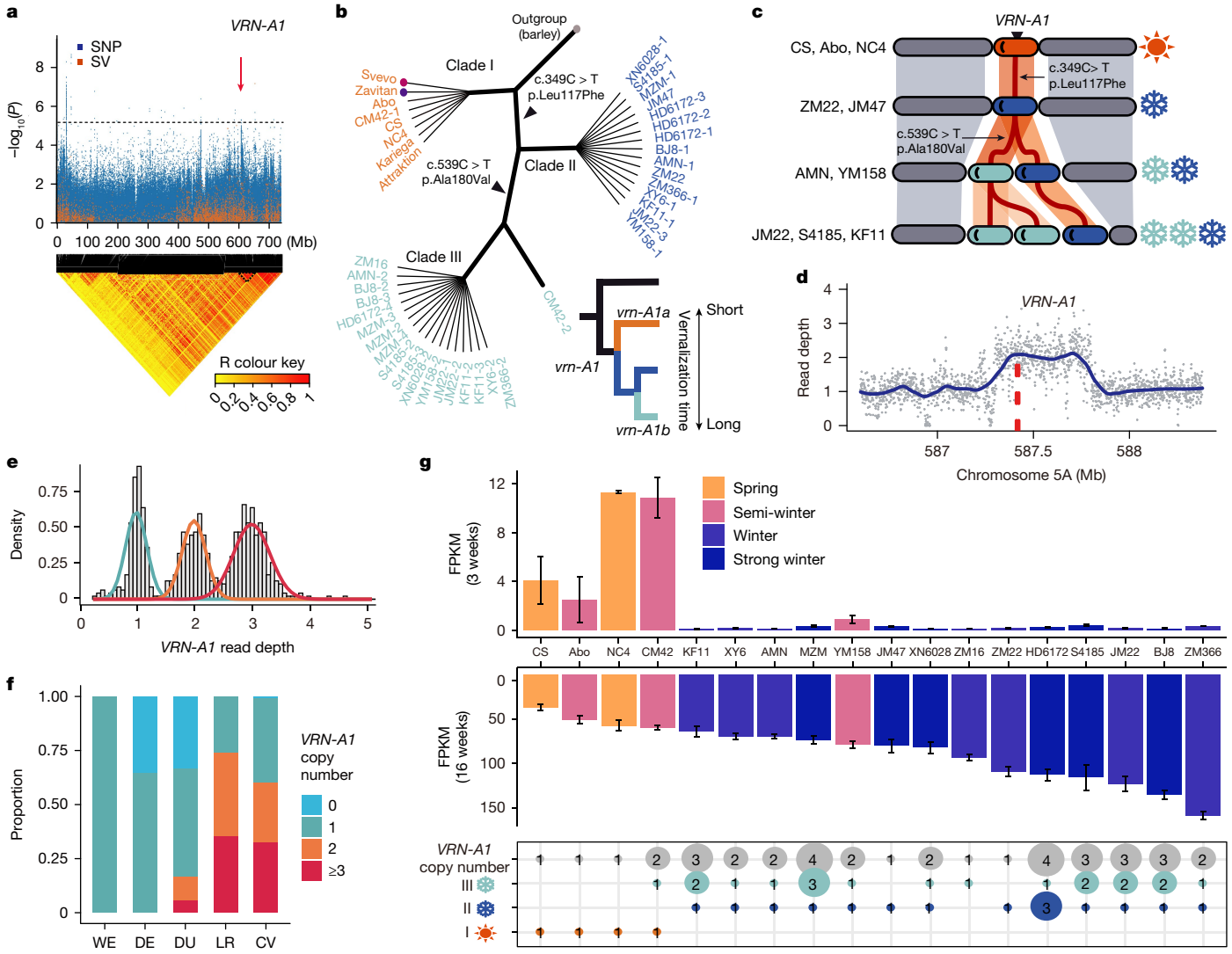


Fig. 3 | Duplication at *VRN-A1* is associated with the winter–spring differentiation of wheat. a, Manhattan plot showing GWAS results in chromosome 5A for heading date. The blue and orange dots represent SNP and SV signals, respectively. The red arrow indicates the *VRN-A1* gene region. **b**, Phylogenetic tree of *VRN-A1* in the 17 newly developed assemblies, Chinese Spring, Kariega, Attraktion, Zavitan, Svevo and Morex genomes. Two SNP variants in the CDS of the *VRN-A1* differentiate into three clades. Clade I contains *VRN-A1* of spring wheat (orange). Clades II and III contain *VRN-A1* of winter wheat (dark blue and light blue, respectively). The phylogeny diagram shows the relationship between three clades corresponding to the different haplotypes of *VRN-A1*. **c**, Collinearity pattern of the *VRN-A1* in Chinese Spring and representative cultivars from the 17 newly developed assemblies. Coloration of the cylindrical fragments represents the *VNR-A1* types in panel **b**. The number of sun and snowflake icons corresponds to the copy number and types of *VRN-A1* duplicates. **d**, The read depth of the *VRN-A1* gene region in YM158. Each point represents an average read depth in 1-kb bin size, and the blue line is a LOESS (locally weighted regression) fit for all dots. The red dashed line represents the *VRN-A1* gene. **e**, The distribution of read depth of *VRN-A1* in common wheat. The green, orange and red lines represent the Gaussian-fitted distribution curves for *VRN-A1* from one to three copies, respectively. **f**, The variation of *VRN-A1* copy number during polyploidization and domestication in wheat. CV, hexaploid wheat cultivar; DE, domesticated emmer wheat; DU, durum wheat; LR, hexaploid wheat landrace; WE, wild emmer wheat. **g**, The expression pattern of *VRN-A1* in four ecotypes after 3 weeks and 16 weeks of cold treatments. The values on the *y* axis represent the average expression level of *VRN-A1* of three biological replicates ± standard error. FPKM, fragments per kilobase per million mapped reads.

wheat varieties to southeastern China. This observation underscores the relationship between dietary preferences and cultivar selection, indicating that food styles indirectly influence cultivar choice.

Detection of PAVs on 1RS translocations

The 1RS·1BL translocation has gained popularity in Chinese wheat-breeding programmes due to its association with higher yield and good resistance to powdery mildew and yellow rust. Currently, about 45% of cultivars used in production have this translocation. The de novo-assembled 1RS sequences were used to evaluate its diversity and evolution within the wheat genetic background.

Compared with 1BS in the Chinese Spring reference, we identified one synteny lost region (198.6–213.3 Mb) and four high read depth regions at 214.5–215.2 Mb, 234.8–235.5 Mb, 234.8–236.5 Mb and 237.0–239.4 Mb on 1RS pericentromeres. In addition, one 1RS·1BL-associated inversion on 1BL was also detected (Fig. 5a and Supplementary Fig. 18). The resequenced 1RS·1BL translocations were classified into four haplotypes based on these insertions (Fig. 5b).

The comparative analysis of the de novo-assembled 1RS sequences clustered the sequences into three subgroups: Sub1, including ZM22 and S4185 with the longest 1RS assembly; Sub2, including KF11, HD6172 and AMN with a medium size of 1RS assembly; and Sub3, including ZM16, KN9204 and AK58 with the smallest size of 1RS assembly (Extended Data

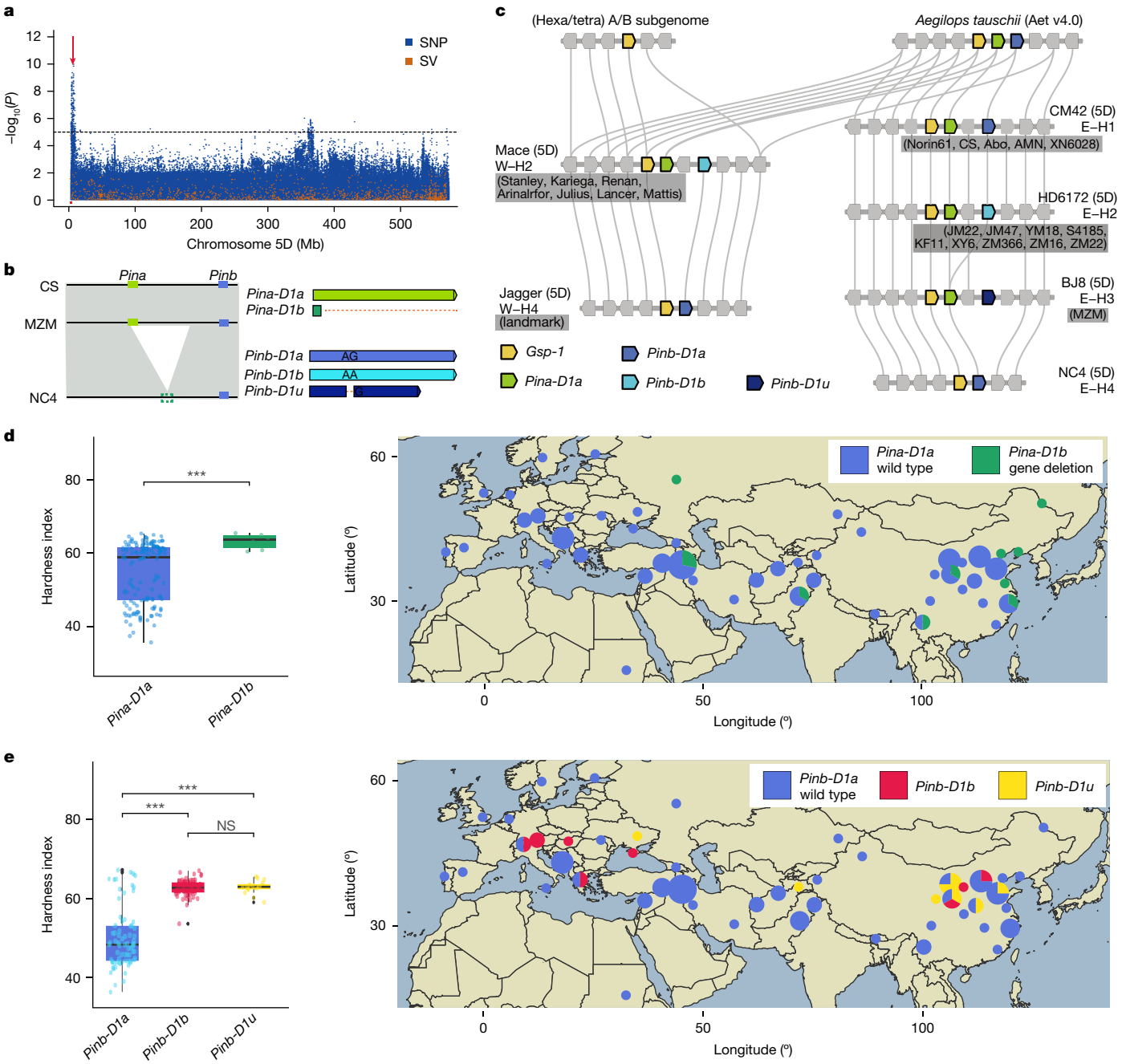


Fig. 4 | Allelic comparison of the *Pina* and *Pinb* genes and their geographical distribution in landrace indicated different priority to the grain hardness in North and South China food culture. a, Manhattan plot of GWAS results for grain hardness in 145 landmark cultivars. Notable association locus is labelled with an arrow and represents *Pina-D1* or *Pinb-D1* genes for further investigation. **b**, *Pina-D1* and *Pinb-D1* alleles found in the 10+ pan-genome and the 17 new assemblies. The left panel displays the collinearity pattern of the *Hardness* (*Ha*) locus, with the green and blue bars indicating *Pina* and *Pinb*, respectively. The right panel demonstrates the transcript structures of these alleles, including *Pina-D1a* (wild type) and *Pina-D1b* (gene deletion) for the *Pina-D1* gene, and *Pinb-D1a* (wild type), *Pinb-D1b* and *Pinb-D1u* for the *Pinb-D1* gene. **c**, Microcollinearity analysis of the *Ha* locus among different assemblies. A filled pentagon represents a coding gene (Supplementary Table 26). The grey lines represent gene orthologue relationships. **d**,**e**, Hapmaps of *Pina-D1* and *Pinb-D1* in landrace (right) and boxplot for grain hardness of different haplotypes of the *Ha* locus (left). Blue represents the wild-type *Pin* gene. Green, red and yellow represent *Pina-D1b*, *Pinb-D1b* and *Pinb-D1u*, respectively. ***$P < 0.001$ and not significant (NS). The map was drawn using the R sf and ggplot2 packages (Data are from resdc (https://www.resdc.cn/data.aspx?DATAID=205)).

Fig. 10a). When using *NOR* as a landmark, the rye subtelomere repeat *pSc200* as probe in combination with *pTa71-2b* in FISH, the PAVs at the 1RS terminal region can be clearly seen by strong signals of *pSc200* among ZM22, ZM16 and AMN. Obvious homozygous and heterozygous partial absences of *pSc200* were detected by FISH with the assistance of *NOR* as a landmark on 1RS in AMN (Fig. 5c). Repeat annotation indicated that the additional 20-Mb fragment mainly contained 1RS-specific subtelomere repeats, and 51–97 genes and homologous fragments were detected in the two newly assembled rye genomes[36,37] (Extended Data Fig. 10a–c). The significant heterogeneity of 1RS reported in Extended Data Fig. 10, along with the well-characterized enhanced yield and favourable root attributes associated with the 1RS segment[38], reasonably contribute to its widespread and persistent use in wheat breeding.

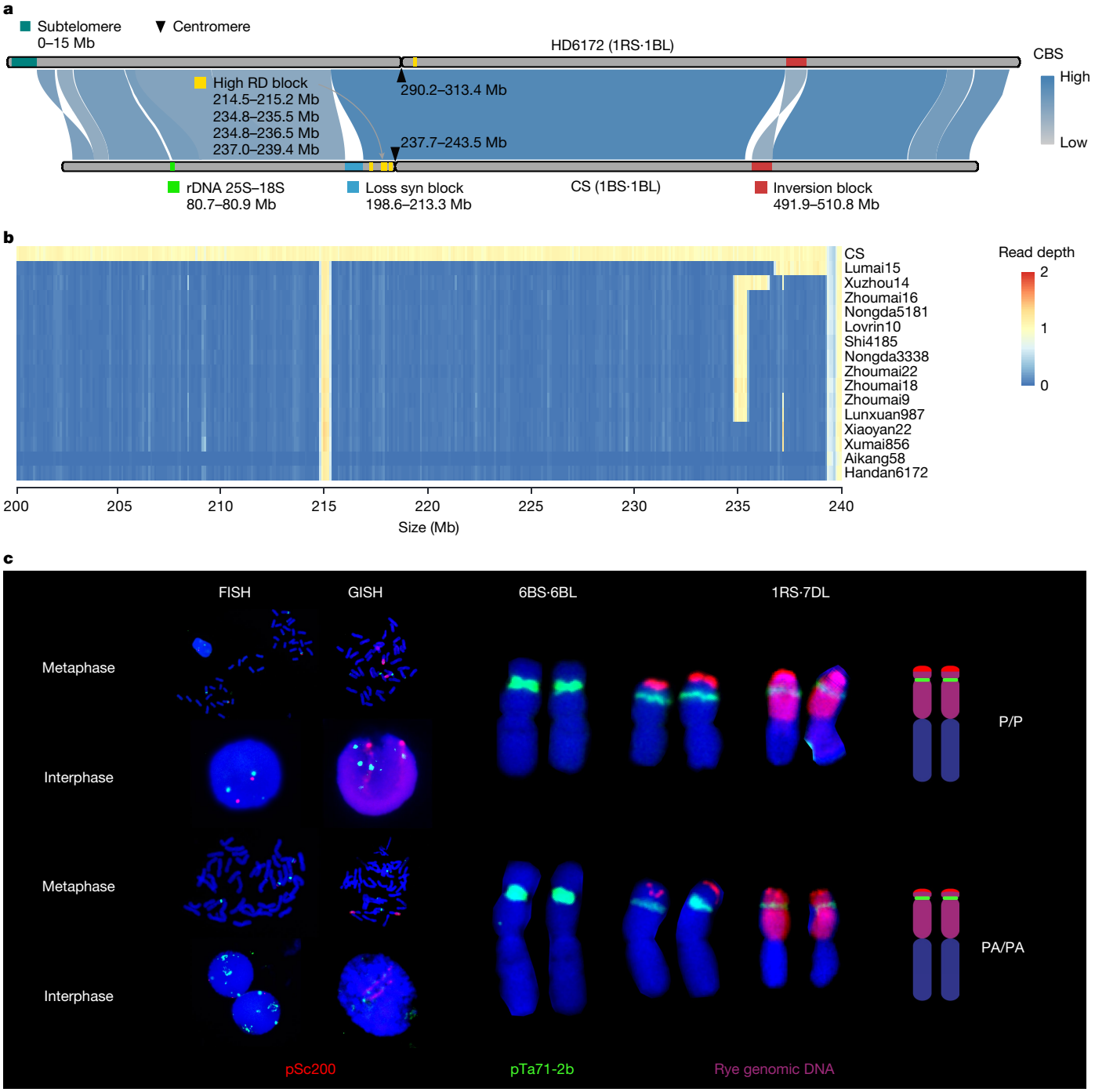


Fig. 5 | Rapid reorganization of 1RS translocated onto wheat chromosomes in the past half century. a, Synteny analysis between the de novo-assembled 1RS·1BL in HD6172 and chromosome 1B in Chinese Spring. The blue bar denotes a synteny (syn) lost region (198.6–213.3 Mb); the yellow bars indicate four high read depth regions (read depth = 1) on 1BS; the red bar represents a 1RS·1BL-associated inversion on 1BL; the dark green bar denotes the subtelomeric sequence of the 1RS; the light green bar represents the rDNA (25S–18S); and the black triangles mark the positions of the centromeres. CBS, collinear block score. **b**, Read depth heatmap across the centromeric regions (200.0–240.0 Mb) on 1RS·1BL cultivars. Four high read depth regions (214.5–215.2 Mb, 234.8–235.5 Mb, 234.8–236.5 Mb and 237.0–239.4 Mb) were detected in the 1RS·1BL cultivars in reference to Chinese Spring chromosome 1B (IWGSC RefSeq v1.1). **c**, Multicolour FISH by DNA repeats detected the big PAV of *pSc200* at the 1RS telomeric region in AMN. The signals of *pSc200* are in red at terminals, *NOR* (*pTa71-2b*) are in green and 1RS are in pink. P, presence; PA, partial absence.

New introgressions from *Ae. tauschii*

CM42 was regarded as the first cultivar bred using CIMMYT synthetic derived from the cross of *T. durum* (AABB) and *Ae. tauschii* (DD). Comparison of CM42 with Chinese Spring revealed two large introgressions on chromosomes 3D and 4D with sizes of 106 Mb (478–584 Mb) and 289 Mb (19–308 Mb), respectively. In addition, several medium-sized (6–19 Mb) introgressions were also detected on chromosome 1D (489–495 Mb), chromosome 2D (37–43 Mb) and chromosome 7D (first 19 Mb; Extended Data Fig. 11). CM42 was not only cultivated as a major cultivar but was also used as a new founder genotype in breeding in southwestern China. This implies that these introgressions did not bring any obvious disadvantage to wheat breeding, and also supports the great potential of *Ae. tauschii* in wheat breeding[39].

Discussion

The significance of pan-genomes lies in the discovery and utilization of important genes present in PAV regions[40]. After its spread to Europe and East Asia, wheat continued to evolve under the two different ecological environments. Genetic introgressions from wild tetraploid wheat led to the accumulation of extensive structure variations in the A and B subgenomes, creating a genomic landscape with distinct geographical features[16]. Breeding crosses between geographical populations facilitated their integration and reshaped the genome landscape of modern wheat cultivars[11].

Across the centromeric region, both DNA and histone proteins were usually heavily methylated and ethylated to maintain the silencing of abundant transposons. This is also considered a major factor in repressing crossover recombination, which has been associated with the presence of special protein at pericentromeres[41,42]. We discovered that the presence of PAVs among cultivars was also associated with reduced crossover recombination in breeding. We believe that frequent introgressions from wild emmer and frequent insertions of younger *CRW* and *Quinta* elements at the core region of centromeres caused abundant PAVs at wheat pericentromeric regions[33,43,44]. This work provides essential information for the selection of parents to optimize the recombination of genes at pericentromeric regions in breeding and gene cloning.

The *VRN-A1* and *PPD-1* genes were regarded as the two most important genes in wheat adaptation to environments[45]. Despite substantial advances in understanding the genetic regulators for wheat vernalization[46,47], the historical switch between spring and winter wheat and the global spread of wheat remains poorly understood. Copy number variation and its effect on expression of *VRN-A1* was, to our knowledge, first observed by Diaz and colleagues, which led to the foundational understanding of the genetic mechanism for wheat to adapt to a broad range of environments[48,49]. However, the duplication unit at *VRN-A1* could not be profiled accurately because of the limitation of genomic resources at that time. We identified copy number variation of *VRN-A1* along with an approximately 0.5-Mb haplotype block among multiple wheat assemblies and decoded the missense mutations accumulated during the evolution of *VRN-A1*. We revealed that the spring-type corresponds to the ancient haplotype, whereas over time, accumulated haplotypes correspond to the winter type (Fig. 3b,c,f). The *VRN-A1* duplications, identified by its three haplotypes, exemplify the high genetic diversity among landraces and modern cultivars in response to changing environmental conditions, especially highlighting global wheat-breeding efforts to maximize yield under the global warming trend. Much higher expression of *VRN-A1* in winter cultivar after full vernalization explains the adaptation of wheat to two-crop season cultivation in warm temperate regions worldwide.

The grain hardness locus is one of the two major loci affecting grain texture and strongly influences end-use processing quality in wheat[50]. The difference in food culture between northern and southern Chinese individuals leads to apparent distinctions in selection and preference. More *Pinb* mutations are retained in ancient landraces in northern China and Europe, which is more pronounced in modern bred varieties. The NC4, Jagger and CDC landmark sharing the *Pina* big deletion mutation closely relate to baking consumption in northwestern China, Europe and America. Therefore, wheat varieties not only serve as commodities but also embody cultural food traditions.

The 1RS·1BL is the most successfully used alien translocations in global wheat improvement. Obvious elimination of the *pSc200* repeat (PAV) at the terminal region of AMN strongly suggested that the 1RS is undergoing rapid evolution in the wheat genome, raising the opportunity for new variations (Fig. 5c). In addition, the discovery of large fragments from ancestral species in CM42 further substantiated the great potential and value of *Ae. tauschii* in wheat breeding.

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/s41586-024-08277-0.

1. International Wheat Genome Sequencing Consortium (IWGSC). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, eaar7191 (2018).
2. Walkowiak, S. et al. Multiple wheat genomes reveal global variation in modern breeding. *Nature* **588**, 277–283 (2020).
3. Salamini, F., Zkan, H., Brandolini, A., Schfer-Pregl, R. & Martin, W. Genetics and geography of wild cereal domestication in the near east. *Nat. Rev. Genet.* **3**, 429–441 (2002).
4. The International Wheat Genome Sequencing Consortium (IWGSC). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* **345**, 1251788 (2014).
5. Feldman, M. & Levy, A. A. Genome evolution due to allopolyploidization in wheat. *Genetics* **192**, 763–774 (2012).
6. Biehl, P. F. et al. Ancient DNA from 8400 year-old catalhöyük wheat: implications for the origin of neolithic agriculture. *PLoS ONE* **11**, e0151974 (2016).
7. Zhao, X. B. et al. Population genomics unravels the Holocene history of bread wheat and its relatives. *Nat. Plants* **9**, 403–419 (2023).
8. Michael F, S. et al. A 3,000-year-old Egyptian emmer wheat genome reveals dispersal and domestication history. *Nat. Plants* **5**, 1120–1128 (2019).
9. Mcclatchie, M. et al. Neolithic farming in north-western Europe: archaeobotanical evidence from Ireland. *J. Archaeol. Sci.* **51**, 206–215 (2014).
10. Liu, X. et al. From ecological opportunism to multi-cropping: mapping food globalisation in prehistory. *Quat. Sci. Rev.* **206**, 21–28 (2019).
11. Hao, C. et al. Resequencing of 145 landmark cultivars reveals asymmetric sub-genome selection and strong founder genotype effects on wheat breeding in China. *Mol. Plant* **13**, 1733–1751 (2020).
12. Zhuang, Q. S. *Chinese Wheat Improvement and Pedigree Analysis* [Chinese] (Agricultural Press, 2003).
13. Murukarthick, J., Mona, S., Nils, S. & Martin, M. Building pan-genome infrastructures for crop plants and their use in association genetics. *DNA Res.* **28**, dsaa030 (2021).
14. Lei, L., Goltsman, E., Goodstein, D., Wu, G. A. & Vogel, J. P. Plant pan-genomics comes of age. *Annu. Rev. Plant Biol.* **72**, 411–435 (2021).
15. Mona, S., Murukarthick, J., Nils, S. & Martin, M. Plant pangenomes for crop improvement, biodiversity and evolution. *Nat. Rev. Genet.* https://doi.org/10.1038/s41576-024-00691-4 (2024).
16. Zhang, X. Y. & Appels, R. in *The Wheat Genome* (eds Appels, R. et al.) 93–111 (Springer, 2023).
17. Castillo, F. A. *The Oxford Handbook of the Archaeology of Diet* (Oxford Univ. Press, 2015).
18. Simon G, K. et al. A putative ABC transporter confers durable resistance to multiple fungal pathogens in wheat. *Science* **323**, 1360–1363 (2009).
19. Fu, D. et al. A kinase-START gene confers temperature-dependent resistance to wheat stripe rust. *Science* **323**, 1357–1360 (2009).
20. Wang, B. et al. De novo genome assembly and analyses of 12 founder inbred lines provide insights into maize heterosis. *Nat. Genet.* **55**, 312–323 (2023).
21. Qin, P. et al. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* **184**, 3542–3558.e16 (2021).
22. Song, L. et al. Reducing brassinosteroid signalling enhances grain yield in semi-dwarf wheat. *Nature* **617**, 118–124 (2023).
23. Németh, A. & Längst, G. Genome organization in and around the nucleolus. *Trends Genet.* **27**, 149–156 (2011).
24. Kishii, M. & Mao, L. Synthetic hexaploid wheat: yesterday, today, and tomorrow. *Engineering* **4**, 552–558 (2018).
25. Guo, W. et al. Origin and adaptation to high altitude of Tibetan semi-wild wheat. *Nat. Commun.* **11**, 5085 (2020).
26. Zhou, Y. et al. *Triticum* population sequencing provides insights into wheat adaptation. *Nat. Genet.* **52**, 1412–1422 (2020).
27. Monat, C., Padmarasu, S., Lux, T., Wicker, T. & Mascher, M. TRITEX: chromosome-scale sequence assembly of Triticeae genomes with open-source tools. *Genome Biol.* **20**, 284 (2019).
28. Athiyannan, N. et al. Long-read genome sequencing of bread wheat facilitates disease resistance gene cloning. *Nat. Genet.* **54**, 227–231 (2022).
29. Kale, S. M. et al. A catalogue of resistance gene homologs and a chromosome-scale reference sequence support resistance gene mapping in winter wheat. *Plant Biotechnol. J.* **20**, 1730–1742 (2022).
30. Li, B. et al. Wheat centromeric retrotransposons: the new ones take a major role in centromeric structure. *Plant J.* **73**, 952–965 (2013).
31. Ahmed, H. I. et al. Einkorn genomics sheds light on history of the oldest domesticated wheat. *Nature* **620**, 830–838 (2023).
32. Wang, Z. et al. Dispersed emergence and protracted domestication of polyploid wheat uncovered by mosaic ancestral haploblock inference. *Nat. Commun.* **13**, 3891 (2022).
33. Cheng, H., Liu, J., Wen, J., Nie, X. & Jiang, Y. Frequent intra- and inter-species introgression shapes the landscape of genetic variation in bread wheat. *Genome Biol.* **20**, 136 (2019).
34. Oliver, S. N., Finnegan, E. J., Dennis, E. S., Peacock, W. J. & Trevaskis, B. Vernalization-induced flowering in cereals is associated with changes in histone methylation at the VERNALIZATION1 gene. *Proc. Natl Acad. Sci. USA* **106**, 8386–8391 (2009).

35. Alonge, M. et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**, 145–161.e23 (2020).
36. Li, G. et al. A high-quality genome assembly highlights rye genomic characteristics and agronomically important genes. *Nat. Genet.* **53**, 574–584 (2021).
37. Rabanus-Wallace, M. T. et al. Chromosome-scale genome assembly provides insights into rye biology, evolution and agronomic potential. *Nat. Genet.* **53**, 564–573 (2021).
38. Gabay, G., Zhang, J., Burguener, G. F., Howell, T. & Dubcovsky, J. Structural rearrangements in wheat (1BS)–rye (1RS) recombinant chromosomes affect gene dosage and root length. *Plant Genome* **14**, e20079 (2021).
39. Zhou, Y. et al. Introgressing the *Aegilops tauschii* genome into wheat as a basis for cereal improvement. *Nat. Plants* **7**, 774–786 (2021).
40. Song, J. M. et al. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nat. Plants* **6**, 34–45 (2020).
41. Saayman, X., Graham, E., Nathan, W. J., Nussenzweig, A. & Esashi, F. Centromeres as universal hotspots of DNA breakage, driving RAD51-mediated recombination during quiescence. *Mol. Cell* **83**, 523–538.e7 (2023).
42. Nambiar, M. & Smith, G. R. Pericentromere-Specific cohesin complex prevents meiotic pericentric DNA double-strand breaks and lethal crossovers. *Mol. Cell* **71**, 540–553.e4 (2018).
43. He, F. et al. Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. *Nat. Genet*. https://doi.org/10.1038/s41588-019-0382-2 (2019).
44. Zhao, J. et al. Centromere repositioning and shifts in wheat evolution. *Plant Commun.* **4**, 100556 (2023).
45. Scott A, B. et al. *Ppd-1* is a key regulator of inflorescence architecture and paired spikelet development in wheat. *Nat. Plants* **1**, 14016 (2015).
46. Yan, L. L. et al. The wheat *VRN2* gene is a flowering repressor down-regulated by vernalization. *Science* **303**, 1640–1644 (2004).
47. Yan, L. et al. Positional cloning of the wheat vernalization gene *VRN1*. *Proc. Natl Acad. Sci. USA* **100**, 6263–6268 (2003).
48. Hazen, S. P. et al. Copy number variation affecting the Photoperiod-B1 and Vernalization-A1 genes is associated with altered flowering time in wheat (*Triticum aestivum*). *PLoS ONE* https://doi.org/10.1371/journal.pone.0033234 (2012).
49. Würschum, T., Boeven, P. H. G., Langer, S. M., Longin, C. F. H. & Leiser, W. L. Multiply to conquer: copy number variations at *Ppd-B1* and *Vrn-A1* facilitate global adaptation in wheat. *BMC Genet.* **16**, 96 (2015).
50. Giroux, M. J. & Morris, C. F. Wheat grain hardness results from highly conserved mutations in the friabilin components puroindoline a and b. *Proc. Natl Acad. Sci. USA* **11**, 6262–6266 (1998).

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2024

Methods

Plant material and genome sequencing

A total of 17 wheat cultivars, including one landraces, 15 modern cultivars and one introduced European cultivar selected based on their historical contribution to wheat breeding and production in China, were used for de novo genome assembly. The high-quality genomic DNA was extracted from seedlings of these 17 varieties using the DNeasy Plant Mini Kit (Qiagen).

PacBio HiFi sequencing

For long-read sequencing, more than 6 μg of sheared genomic DNA was subjected to size selection by the BluePippin system (Sage Science), and approximately 15-kb Sequel Single-Molecule Real-Time (SMRT) bell libraries were prepared according to the manufacturer's instructions (Pacific Biosciences). The libraries were sequenced on 16–20 SMRT cells using the PacBio Sequel II platform in circular consensus sequencing mode with 30-h movie (Supplementary Table 2).

Hi-C sequencing

For Hi-C sequencing, the high-molecular-weight genomic DNA of the 17 cultivars was extracted. Plant material fixation, nuclei extraction, DNA crosslinking and restriction enzyme ligation were performed as previously described[51]. In brief, the digested DNA was blunt-ended and incorporated with biotin-14-dCTP (Invitrogen) before ligation with T4 DNA ligase at room temperature for 4 h. After purification, DNA was sheared by sonication using Covaris S220. Subsequently, end-repaired DNA was separated, purified and ligated with adapters for library preparation. The final libraries for Hi-C sequencing were constructed using the DpnII restriction endonuclease and were sequenced in paired-end mode (2 × 150 bp) using Illumina NovaSeq 6000 (Supplementary Table 2).

De novo genome assembly and evaluation

The PacBio HiFi reads were assembled using Hifiasm (v0.13-r308)[52] with default parameters to generate draft contigs for each cultivar. The assembled contigs for each cultivar were used as a reference to map the Hi-C data for all cultivars using HiCUP (v0.7.3-1)[53]. Only the uniquely mapped and valid di-tags paired-end reads were considered high quality and mapped to the respective contigs of each cultivar to obtain aligned BAM files. The contigs were then clustered using ALLHiC[54] ('ALLHiC_partition -e GATC -k 21'). Finally, the resulting draft assemblies were manually corrected using Juicebox (v1.11.08)[55]. To evaluate the completeness of the assembled genomes, we used the Benchmarking Universal Single-Copy Orthologs (BUSCO)[56] tool, using the embryophyta_odb10 gene set as a reference. Furthermore, the LAI[57] was used to evaluate assembly continuity by full-length long-terminal repeats retrotransposons.

Transcriptomic sequencing

The total RNA of four cultivars (AMN, XY6, ZM16 and JM22) from eight tissues, including leaves, roots and stems at the seedling stage (three-leaf stage), leaves, roots and stems at the tillering stage and seeds at 10 and 20 days post-anthesis, was extracted with a TRIzol reagent (Invitrogen). RNA quality was estimated using agarose gel electrophoresis, Nanodrop, Qubit 2.0 and Agilent 2100 Bioanalyzer. The resulting libraries were sequenced in paired-end mode (PE150) using the DNBSEQ-T7 system at Smart Genomics Technology, which produced around 10 Gb of data for each tissue in each sample.

For PacBio Iso-seq, RNA from eight sets of tissues (same as the above) were extracted at the respective time points (Supplementary Table 4). The PacBio Iso-seq libraries were constructed for AMN, XY6, ZM16 and JM22, with an insert size of 0–5 kb, followed by sequencing on the PacBio Sequel II platform to generate more than 10 Gb of data per library (Supplementary Table 5). The sequencing data were further processed using SMRTlink (v9.0.0; https://www.pacb.com/support/software%20downloads/).

NLR family identification

NLR-Annotator[58] (https://github.com/steuernb/NLR-Annotator) was used for annotating NLR genes in all assemblies with default parameters. The NLR sequences that contain the NBARC and LRR domains were further classified into five types as previously described[2], based on the domain features, namely, CC-NBARC-LRR and NBARC-LRR.

Repeat annotation

For repeat annotation of the 17 newly assembled genomes, a repeat library combined from ClariTeRep (https://github.com/jdaron/CLARI-TE) and TREP-DB (http://botserv2.uzh.ch/kelldata/trep-db/downloads/trep-db_complete_Rel-16.fasta.gz) was constructed and used to annotate repeats with RepeatMasker (v4.0.5)[59] (http://repeatmasker.org/). In addition, the tandem repeats were annotated by Tandem Repeats Finder (v409)[60] (Supplementary Table 8 and Supplementary Fig. 2).

De novo gene structure annotation for AMN, XY6, ZM16 and JM22

From each assembled genome, gene models were predicted by integrating homology-based prediction, ab initio prediction and transcriptome evidence. Protein sequences from *Triticum aestivum*, *Triticum urartu*, *Triticum turgidum*, *Aegilops tauschii*, *Hordeum vulgare*, *Brachypodium distachyon*, *Oryza sativa* and *Zea mays* were aligned to each genome using Wublast (v2.0) with an *E*-value cut-off of 1×10^{-5}, and the hits were joined using SOLAR (v0.9.6)[61]. GeneWise (v2.4.1)[62] was used to predict the gene structure for each Wublast hit. The gene structures predicted by GeneWise were denoted as Homo-set (homology-based prediction gene set). The RNA sequencing data of the four cultivars (AMN, XY6, ZM16 and JM22) were aligned to the genomes using PASA (v2.4.1)[63], and the resulting predictions were denoted as PASA-T-set, which was used as the training data for ab initio gene prediction. Five ab initio gene prediction programs, including Augustus (v2.5.5)[64], Genscan (v1.0)[65], Geneid[66], GlimmerHMM (v3.0.1)[67] and SNAP[68], were used to predict coding regions in the repeat-masked genome. The RNA sequencing data were mapped to the assemblies using TopHat (v2.0.8)[69], and the transcripts (denoted as 'Cufflinks-set') were assembled using Cufflinks (v2.1.1)[70]. RNA sequencing data from the Illumina platform were also assembled using Trinity (v2.1.1), creating pseudo-expressed sequence tags. These pseudo-expressed sequence tags were mapped to the assembly, and gene models (PASA-set) were predicted using PASA (v2.2.0). RNA isoform sequencing data were directly mapped to the genome using BLAT and assembled using PASA. Gene models created using PASA were denoted as the PASA-ISO set. This set was also used to train the model for ab initio gene prediction. Gene model evidence from the Homo-set, PASA-T-set, Cufflinks-set and ab initio programs were combined by EvidenceModeler[71] to obtain a non-redundant set of gene annotations. Priorities for each type of evidence were set as follows: PASA-T-set > Homo-set > Cufflinks-set > Augustus > GeneID = SNAP = GlimmerHMM = Genscan.

These gene models were further classified into high-confidence and low-confidence protein-coding genes based on a stringent confidence classification method. BLASTP[72] was used to align the predicted proteins to known protein datasets (*T. turgidum*, *T. aestivum*, *B. distachyon*, *H. vulgare*, *T. urartu*, *Ae. tauschii*, *O. sativa* and *Z. mays*) with an *E*-value cut-off of 1×10^{-10}. For each gene model, we selected the best-matching reference protein as a template sequence and defined the transcript sequence with maximum coverage of the template as the representative sequence. Genes were designated high confidence when there was a significant BLAST hit to reference proteins, and the similarity between representative protein sequence and the respective template sequence was above the threshold in at least two species (more than 70% for

O. sativa, *Z. mays* and *B. distachyon*; and more than 90% for *H. vulgare*, *T. urartu*, *Ae. tauschii*, *T. turgidum* and *T. aestivum*).

The gene functions were predicted by searching protein sequences in five protein/function databases, including NCBI-NR[73], InterPro[74], Gene Ontology, the KEGG[75] and Swiss-Prot[76]. InterPro and Gene Ontology databases were queried using InterProScan[77] with parameters: '-f TSV -dp -goterms -iprlookup'. The NCBI-NR, Swiss-Prot and KEGG databases were searched using BLAST with an *E*-value cut-off of 1×10^{-5} (Supplementary Table 9).

Standardization of gene structure annotation and construction of the pan-genome gene set

For the gene structure annotation of another 14 genomes and normalization of the differences in annotation results caused by different annotation methods, we adopted a method of first constructing a pan-genome set and then mapping the pan-genome set onto the genomes to reannotate all the genomes. We used a stepwise strategy to construct the pan-genome set, using the gene annotations from seven genomes, including AMN, XY6, ZM16, JM22 and the high confidence genes of previously published Chinese Spring, Fielder and Zang1817. Pairwise collinearity analysis was performed for the seven assemblies using JCVI[78] with default parameters. The gene annotation of Chinese Spring was used as the initial gene set, and genes from the other six assemblies were added in a stepwise manner. A gene was added to the gene set and assigned a new locus ID when no collinear gene was found in the existing gene set. The resulting pan-genome set was mapped to each genome assembly using gmap[79]. The alignment results were filtered with a relatively strict cut-off: identity of more than 90% and query coverage of more than 90%. Finally, the number of standardization annotated genes for each assembly ranges from 148,999 to 154,828 (Supplementary Tables 10 and 12). BUSCO evaluation showed that the annotation results after standardization were better than those before standardization (Supplementary Table 11).

Variant calling

For identifying variations, the high-quality paired-end reads were mapped to the Chinese Spring genome (IWGSC RefSeq v2.1) using BWA-MEM (v0.7.17-r1188)[80] with parameters '-t 4 -k 32 -M'. To reduce duplicates generated by PCR amplification before sequencing, duplicated reads were removed using SAMtools (v0.1.1)[81]. The de-duplicated reads were used to call the genomic variants for each line (in GVCF format) using the Sentieon DNAseq software package[82] and merged together. To obtain high-quality variants, the raw variants were processed using GATK[83] to retain variants with 'QD > 2.0, FS < 60.0, MQ > 20.0, MQRankSum > −12.5 and ReadPosRankSum > −8.0'. Next, the following potential low-quality variants were removed with the following four parameters: (1) two alleles only, (2) missing rate ≤ 0.2, (3) genotype quality for individual ≥ 5, and (4) minor allele frequency ≥ 0.05. Consequently, a total of 37,466,916 SNPs and 2,892,919 insertions and deletions (InDels) were retained. The identified SNPs and indels were annotated with ANNOVAR (v2013-05-20)[84] and divided into the following groups: variations occurring in intergenic regions within 1 kb upstream (downstream) of transcription start (stop) sites, in coding sequences and introns, based on genome annotation information (Supplementary Tables 17 and 18).

Identification of SVs

All the 17 new and three previously published assemblies were aligned to the Chinese Spring genome (IWGSC RefSeq v2.1) using MUMmer (v4.0)[85] with parameters: '--maxgap=500 --mincluster=1000'. Next, the alignments were filtered using delta-filter with parameters: '-i 95 -l 5000 -1'. The aligned results were processed using 'show-coords -THrd' to obtain the coordinates files, which were used to detect SVs using the SyRI (v1.6)[86] pipeline with default parameters. The SyRI outputs were then converted into three types of SVs: PAVs, inversions and translocations, according to the previously published methods[21]. The CPL, DEL, DUP/INVDP (loss) variants and the sequences in HDR, NOTAL and TDM were grouped as absence SVs (relative to Chinese Spring) in a given accession. The CPG, INS and DUP/INVDP (gain) variants, and the query sequences in HDR, NOTAL and TDM were classified as presence SVs (relative to Chinese Spring) in a given accession. The INV variants were regarded as inversion SVs relative to Chinese Spring; the TRANS and INVTR were both regarded as translocation SVs relative to Chinese Spring in each accession. For details about these definitions, refer to https://schneebergerlab.github.io/syri/fileformat.html. All of the SVs detected by SyRI were merged using SURVIVOR (v1.0.7)[87] with default parameters, which were genotyped for the 144 accessions using svtyper[88] (Supplementary Table 19).

Validation of PAVs and translocations

A total of 60 PAVs were randomly selected from PAVs between the CM42 and Chinese Spring. The long reads from CM42 were aligned to the Chinese Spring genome to check the deletions. The short reads from Chinese Spring were aligned to CM42 to check presences. The alignments were manually inspected using the Integrative Genomics Viewer. Furthermore, we also manually validated a subset of translocation events by aligning the Hi-C data from these samples to both its own genome and the Chinese Spring genome. When aligning the Hi-C data for MZM and BJ8 to their own genome, we observed lower inter-chromosomal interaction strength than intra-chromosomal interaction strength within MZM and BJ8 chromosomes. However, when aligning their Hi-C data to the Chinese Spring genome, we discovered strong interaction signals between chromosome 4A and chromosome 1D, indicating a translocation between MZM or BJ8 with Chinese Spring in that genomic region. We applied the same method to demonstrate a translocation between AMN and Chinese Spring in the genomic region between chromosome 7D and chromosome 1B, that is, 1RS·7DL and 7DS·1BL (Supplementary Fig. 7).

Diversity analysis of assembled varieties

The variant call format (VCF) files from three previous studies[11,26] were retrieved, combined and filtered to retain hexaploid accessions and polymorphisms detected in three studies. The 10+ project sequencing data[2] and synthetic next-generation sequencing reads of KF11 generated by ART (v4.17.16)[89] were aligned to Chinese Spring RefSeq (v2.1) using the BWA (v0.7.15-r1140) software. Alignment files from the accessions assembled here were then used for variant calling by GATK (v.3.8). The variant files from the whole-genome sequencing studies, 10X Genomics and KF11 were then merged and subjected to *t*-distributed stochastic neighbour embedding (t-SNE) using Rtsne (v0.17)[90] in R (v.3.6.1). Correspondingly, the whole-genome scale GGNet of 145 landmark cultivars was constructed by ggComp[91].

Analysis of PAVs and recombination rate density distribution in the centromeric region

To investigate the effect of PAVs on recombination rates, we utilized the FastEPRR package[92] to analyse SNP data obtained from resequencing 145 wheat varieties. We calculated the recombination rates on each chromosome within 1-Mb windows and counted the number of indels within each window. Furthermore, we aligned previously reported[30] Gypsy transposons RLG_famc8.3 (*Cereba*, also known as *CRW*), RLG_famc8.1 (Quinta-1) and RLG_famc8.2 (Quinta-2), known for their specificity to centromere regions, to the Chinese Spring (v2.1) genome and calculated their density distribution across chromosomes. Regions enriched with such transposons were defined as centromere regions of Chinese Spring v2.1 genome.

Identification of SV hotspot regions

We calculated the distribution of SV breakpoints for each 1-Mb window (with a step size of 500 kb) along each chromosome. Next, all 1-Mb

windows were ranked in descending order according to the numbers of SVs within the window. We defined the top 10% and top 10–20% of all windows with the highest frequency of SV breakpoints as SV hotspots, then merged all of the continuous hotspot windows as the 'hotspot regions' (Extended Data Fig. 5d).

Selective signals of adaptation

To identify potential selective signals during different genetic improvement time periods, a sliding-window approach (window size = 5 Mb and sliding step size = 2 Mb) was applied based on both SNPs and SVs to quantify the levels of genetic differentiation (F_{ST}) between 1950s and 1980s or 2000s using VCFtools (v0.1.14)[93].

GWAS

We performed GWAS with SNPs and SVs for two agronomic traits — ecotype and grain hardness index — which were measured by a single-kernel characterization system for grains of 145 resequenced landmark cultivars harvested in 2022 at Xinxiang and Beijing. Traits including heading date, flowering date, plant height, effective tiller number, grain number per head and thousand-grain weight were also used for GWAS analysis using the EMMAX package[94]. The kinship matrix of pairwise genetic similarities, which were derived from the simple matching coefficients, was used as the variance–covariance matrix of the random effects and was also calculated by EMMAX. We then used a threshold of $P < 1.0 \times 10^{-5}$ to identify significant association signals for subsequent analysis. Significant GWAS signals associated with multiple traits were detected at the *VRN-A1* locus on chromosome 5A and the *Ha* locus on chromosome 5D.

Phylogenetic analysis

Multiple sequence alignments of *VRN-A1* CDS sequences in 17 newly developed assemblies, Zavitan, Svevo and barley were performed using MAFFT (v7.471)[95]. The phylogenetic tree was built from the MAFFT alignment using the FastTree (v2.1.11)[96] tool with the FFT-NS-2 model to generate maximum likelihood trees.

Gene copy number variation analysis

To validate the gene copy number variations of *VRN-A1* based on the resequenced data, we compared the copy number from annotation based de novo assemblies and the copy number calculated by the average read depth (ARD) in each window via the 'coverage' function of bedtools (v2.26.0)[97]. High consistency was detected between the de novo assemblies and resequenced data for ZM16, XY6, AMN, JM22 and ZM22. The ARDs were then normalized by dividing the mode value of the total ARDs in each gene. Gaussian mixture modelling of the ARDs of *VRN-A1* was performed on the normalized ARDs of 423 accessions using mclust[98] to infer the copy number of *VRN-A1*.

Synteny analysis

To identify syntenic gene blocks between two assemblies, all-vs-all BLASTP ($E < 1 \times 10^{-5}$, top five matches) was performed for the high-confidence gene sets of each genome pair. GeneTribe[99] was used to define syntenic blocks based on the presence of at least five synteny gene pairs, as well as to conduct homology inference with default settings. All-vs-all homology inferences were performed for diploid sub-assemblies. Collinearity analysis of the *VRN-A1* and PIN gene regions based on sequences was also carried out using GeneTribe, with parameters: '-r True -c False -e 1e-5 -b 75'.

Winter and spring habit verification

All of the 17 cultivars used in the study were sowed in three months (27 February, 23 March and 11 April 2023) following the national winter and spring habit verification standard in the Luoyang Academy of Agriculture and Forestry, Henan province, which was assigned by the Ministry of Agriculture and Rural Affairs for verification of winter and spring characters of new wheat cultivars.

Development of KASP and indel markers

SNPs between *Pinb-D1a* and *Pinb-D1b* were used to develop molecular markers for analysis. SNP markers were converted to Kompetitive allele-specific PCR (KASP) markers using Polymarker (http://www.polymarker.info/). Primer pairs for *Pina-D1a* (wild type) and *Pina-D1b* (gene deletion) were designed using the PrimerServer tool of WheatOmics[100]. Information on newly developed markers is listed in Supplementary Table 28.

Collinearity analysis and repeat sequence annotation for 1RS

Using a custom repeat database, the short-arm terminal extension sequences of 1RS translocation cultivars were annotated and soft-masked with RepeatMasker (v4.0.7; http://www.repeatmasker.org). Tandem repeats were identified with Tandem Repeat Finder (v4.09)[60] with the parameters: '2 5 7 80 10 50 2000 -l 10 -f -d -m -ngs'. Synteny plots between the short-arm terminal extension sequences of 1RS translocation cultivars without gene-based collinearity were conducted using GenomeSyn (v1.41)[101]. With GenomeSyn, the extension sequences were truncated into 500-kb fragments for subsequent collinearity analysis. We used minimap2 (v2.22-r1110)[102] with default parameters to perform alignment of the sequences.

FISH and GISH of the 1RS translocations

To determine the specific location of the centromere in each genome, we conducted a BLAST search on each genome assembly of the 1RS translocation cultivars using centromere-specific LTR retrotransposons (Cereba and Quinta) with *E*-value cut-off of 1×10^{-5}. The obtained 1RS sequences from each genome were subsequently used for further analysis. To establish the evolutionary relationship of the 1RS in different cultivars, we aligned all single-copy gene protein sequences using MUSCLE (http://www.drive5.com/muscle/) and combined all of the alignment results to create a super alignment matrix of 1RS. Subsequently, we constructed the phylogenetic tree with the maximum likelihood method implemented in RAxML (http://sco.h-its.org/exelixis/web/software/raxml/index.html) with 1,000 bootstrap replicates. The SVs among 1RS were identified using ZM22 as the reference genome (with the largest 1RS assembly) following previously described methods using MUMmer and SyRI. Metaphase chromosomes for FISH and GISH analyses were prepared as previously described[103]. The 1RS chromosome arms in AMN, ZM16 and ZM22 were identified using the genomic DNA of rye in conjunction with the synthetic oligonucleotide probes of *NOR* in green colour and pSc200 in red colour, a rye terminal-specific repeats[104,105]. Images were captured by a ZEISS Imager Z2 microscope and were subsequently processed with Adobe Illustrator to highlight key features.

Reporting summary

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

Data availability

All data are available in this paper, the Supplementary Information or at publicly accessible repositories. The data in the public repositories include all raw reads and assembled sequence data (Supplementary Table 29) for wheat pan-genomics in the BIG Data Center under BioProject ID PRJCA021345. All materials are available from X.Z. on request. Although DNA samples of the 17 assembled genotypes are freely available, the seeds of these genotypes can be obtained following China Legislation on Crop Seeds and Material Transfer Agreement. There is no concern for researchers in China to access these seeds.

Code availability

The source code and scripts used in the paper have been deposited in GitHub (https://github.com/Xiaoming8102/WheatPangenome).

51. Xie, T. et al. *De novo* plant genome assembly based on chromatin interactions: a case study of *Arabidopsis thaliana*. *Mol. Plant* **8**, 489–492 (2015).
52. 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).
53. Wingett, S. et al. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res.* **4**, 1310 (2015).
54. Zhang, J. et al. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nat. Genet.* **50**, 1565–1573 (2018).
55. Burton, J. N. et al. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
56. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
57. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR assembly index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
58. Burkhard, S. et al. The NLR-Annotator tool enables annotation of the intracellular immune receptor repertoire. *Plant Physiol.* **183**, 468–482 (2020).
59. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* https://doi.org/10.1002/0471250953.bi0410s05 (2009).
60. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
61. Yu, X. J., Zheng, H. K., Wang, J., Wang, W. & Su, B. Detecting lineage-specific adaptive evolution of brain-expressed genes in human using rhesus macaque as outgroup. *Genomics* **88**, 745–751 (2006).
62. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
63. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
64. Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
65. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
66. Guigo, R. Assembling genes from predicted exons in linear time with dynamic programming. *J. Comput. Biol.* **5**, 681–702 (1998).
67. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).
68. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
69. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
70. Ghosh, S. & Chan, C. K. Analysis of RNA-seq data using TopHat and Cufflinks. *Methods Mol. Biol.* **1374**, 339–361 (2016).
71. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
72. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
73. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61–D65 (2007).
74. Hunter, S. et al. InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**, D211–D215 (2009).
75. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
76. Boeckmann, B. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
77. Quevillon, E. et al. InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116–W120 (2005).
78. Tang, H. et al. Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
79. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
80. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
81. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
82. Weber, J. A., Aldana, R., Gallagher, B. D. & Edwards, J. S. Sentieon DNA pipeline for variant detection-Software-only solution, over 20× faster than GATK 3.3 with identical results. *PeerJ PrePrints* **4**, e1672v1672 (2016).
83. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
84. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
85. Marcais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
86. Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277 (2019).
87. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
88. Chiang, C. et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods* **12**, 966–968 (2015).
89. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).
90. Laurens, V. D. M. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).
91. Yang, Z. et al. ggComp enables dissection of germplasm resources and construction of a multiscale germplasm network in wheat. *Plant Physiol.* **188**, 1950–1965 (2022).
92. Gao, F., Ming, C., Hu, W. & Li, H. New software for the fast estimation of population recombination rates (FastEPRR) in the genomic era. *G3* **6**, 1563–1571 (2016).
93. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
94. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
95. Katoh, K., Asimenos, G. & Toh, H. Multiple alignment of DNA sequences with MAFFT. *Methods Mol. Biol.* **537**, 39–64 (2009).
96. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
97. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
98. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R J.* **8**, 289–317 (2016).
99. 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).
100. Ma, S. et al. WheatOmics: a platform combining multiple omics data to accelerate functional genomics studies in wheat. *Mol. Plant* **14**, 1965–1968 (2021).
101. He, W. et al. NGenomeSyn: an easy-to-use and flexible tool for publication-ready visualization of syntenic relationships across multiple genomes. *Bioinformatics* **39**, btad121 (2023).
102. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
103. Han, F., Lamb, J. C. & Birchler, J. A. High frequency of centromere inactivation resulting in stable dicentric chromosomes of maize. *Proc. Natl Acad. Sci. USA* **103**, 3238–3243 (2006).
104. Fu, S., Chen, L., Wang, Y., Li, M. & Tang, Z. Oligonucleotide probes for ND-FISH analysis to identify rye and wheat chromosomes. *Sci. Rep.* **5**, 10552 (2015).
105. Tang, Z., Yang, Z. & Fu, S. Oligonucleotides replacing the roles of repetitive sequences pAs1, pSc119.2, pTa-535, pTa71, CCS1, and pAWRC.1 for FISH analysis. *J. Appl. Genet.* **55**, 313–318 (2014).

Acknowledgements We appreciate Z. F. Lu for discussion on the *VRN-A1* expression work and W. X. Wang for help on bioinformatics analysis. This project was funded by the National Key Research and Development Program of China (2023YFD1000400 and 2022YFD1201503). This project was also funded by the National Natural Science Foundation of China (grant no. 32322059) and the Innovation Program of Chinese Academy of Agricultural Sciences (CAAS-CSCB-202401). R.K.V. thanks Food Futures Institute, Murdoch University and Grains Research & Development Corporation (project nos. UMU2404-003RTX and WSU2303-001RTX) for supporting this work in part.

Author contributions X.Z. together with W.G. designed the project. X.Z., W.G. and R.K.V. supervised the execution and completion of the project. C.J., X.X. and L.C. performed the bioinformatics analysis. C.H managed the fieldwork and prepared the samples. L.Z. conducted the cytogenetic experiments. V.G., Z.W., Y.Z., T.L., J.F., A.C., J.H., H.L., G.D., X.L., J.J., L.M. and X.W. contributed to the conducting of experiments, data analysis and interpretation for various sections of the paper. X.Z., W.G., Y.X., V.G. and R.A. wrote the paper, with input from all authors. All authors read and approved the final manuscript.

Competing interests The authors declare no competing interests.

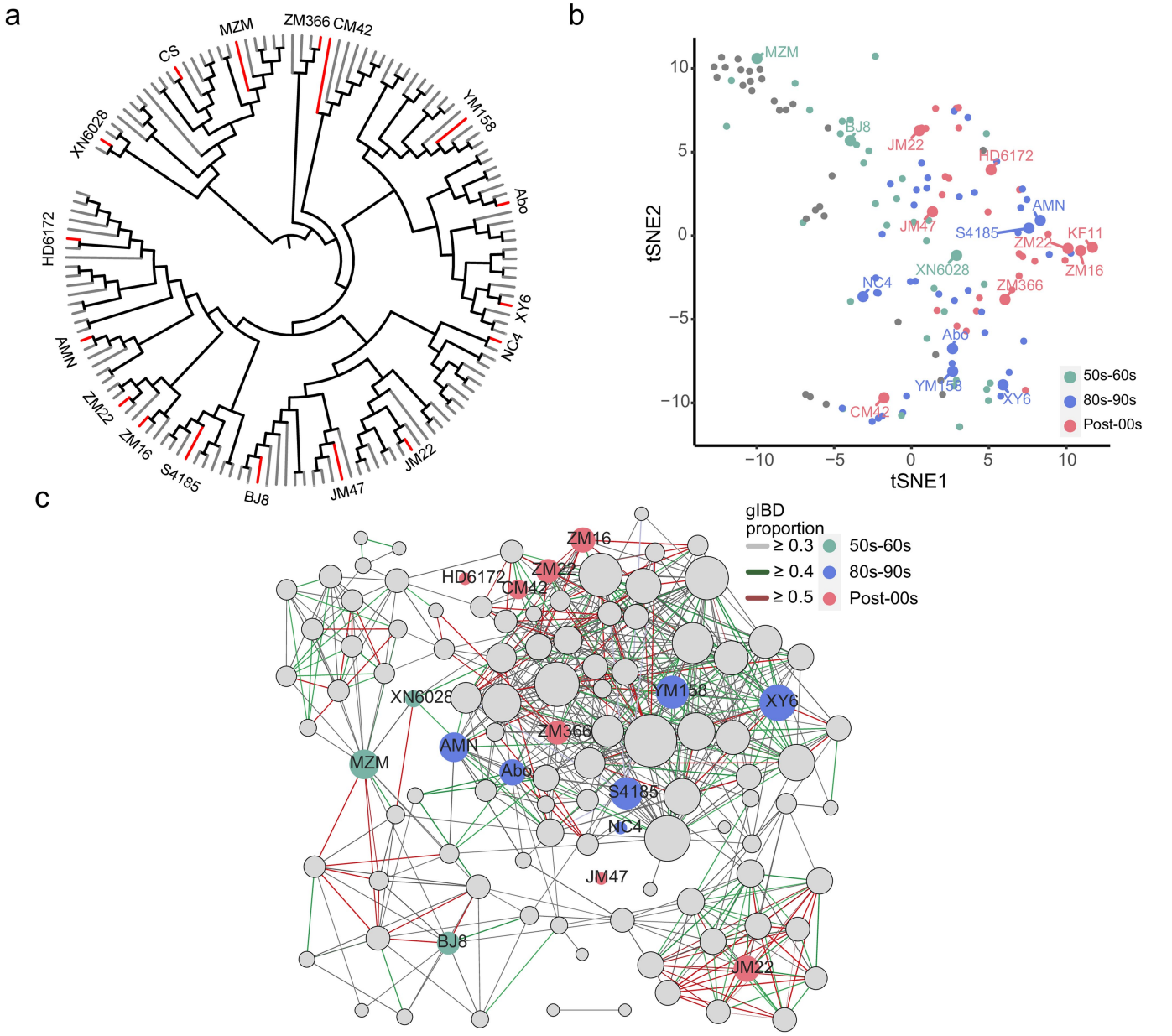
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-08277-0.

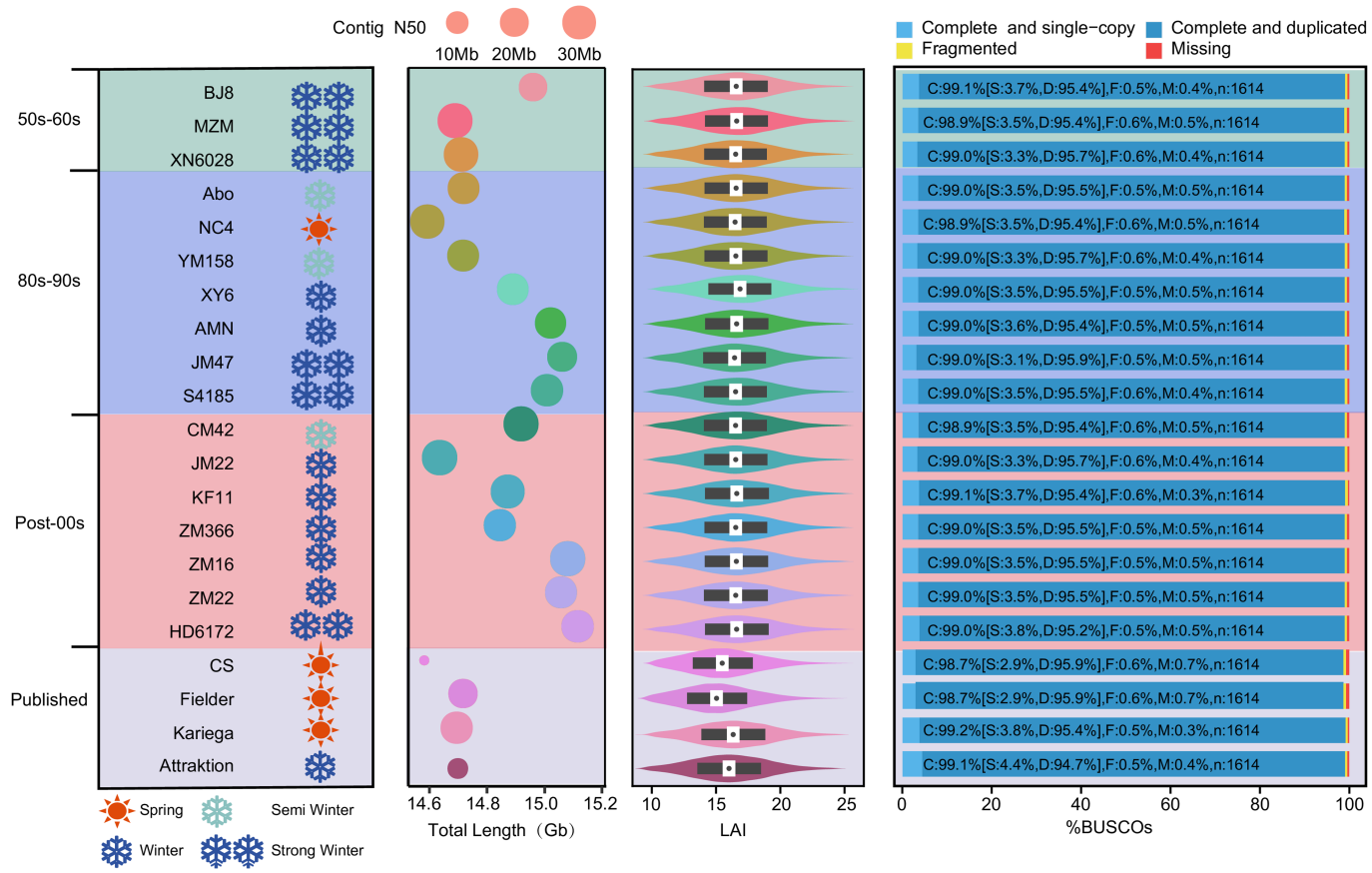
Correspondence and requests for materials should be addressed to Rajeev K. Varshney, Weilong Guo or Xueyong Zhang.

Peer review information *Nature* thanks Erik Garrison, Sean Walkowiak and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

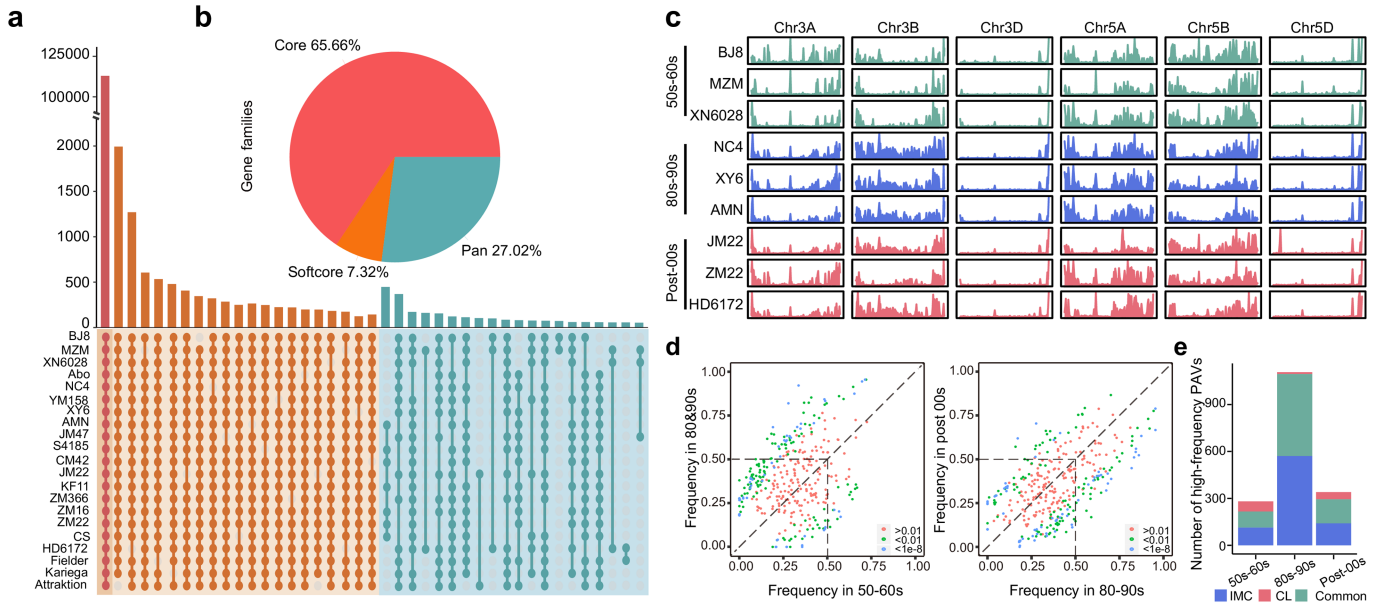
Reprints and permissions information is available at http://www.nature.com/reprints.



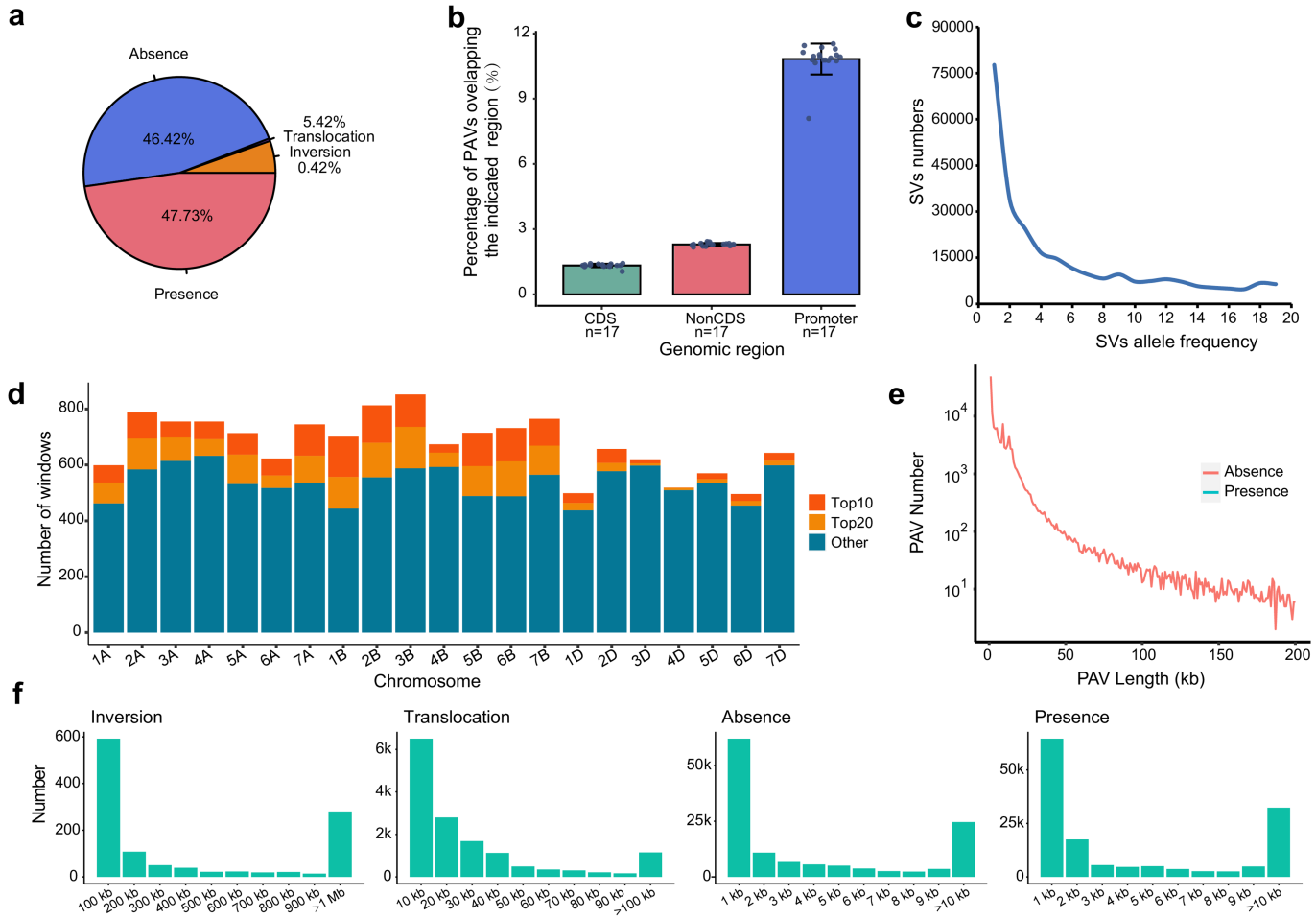
Extended Data Fig. 1 | The genome diversity and representation of 17 assembled genomes for the 145 resequenced landmark cultivars (Hao et al.[11]). **a**, Phylogenetic tree of all accessions inferred from whole-genome SNPs. The red lines indicate the cultivars used for de novo genome assembly. **b**, The t-distributed stochastic neighbor embedding (t-SNE) analysis based on the SNPs revealed the genetic relationships between the 17 de novo assemblies and the 145 landmark cultivars. **c**, The genome wide GGNet of 145 landmark cultivars. The 17 de novo assembled cultivars were marked in green for 1950s, blue for 1980s-1990s and red for post 2000s. The edge colors indicate the ranges of the gIBD ratio (genome similarity) for accession pairs. Only the edges in which the gIBD ratio ≥ 30% are shown. Grey edges, 40% > gIBD ratio ≥ 30%; Green edges, 50% > gIBD ratio ≥ 40%; Red edges, gIBD ratio ≥ 50%.



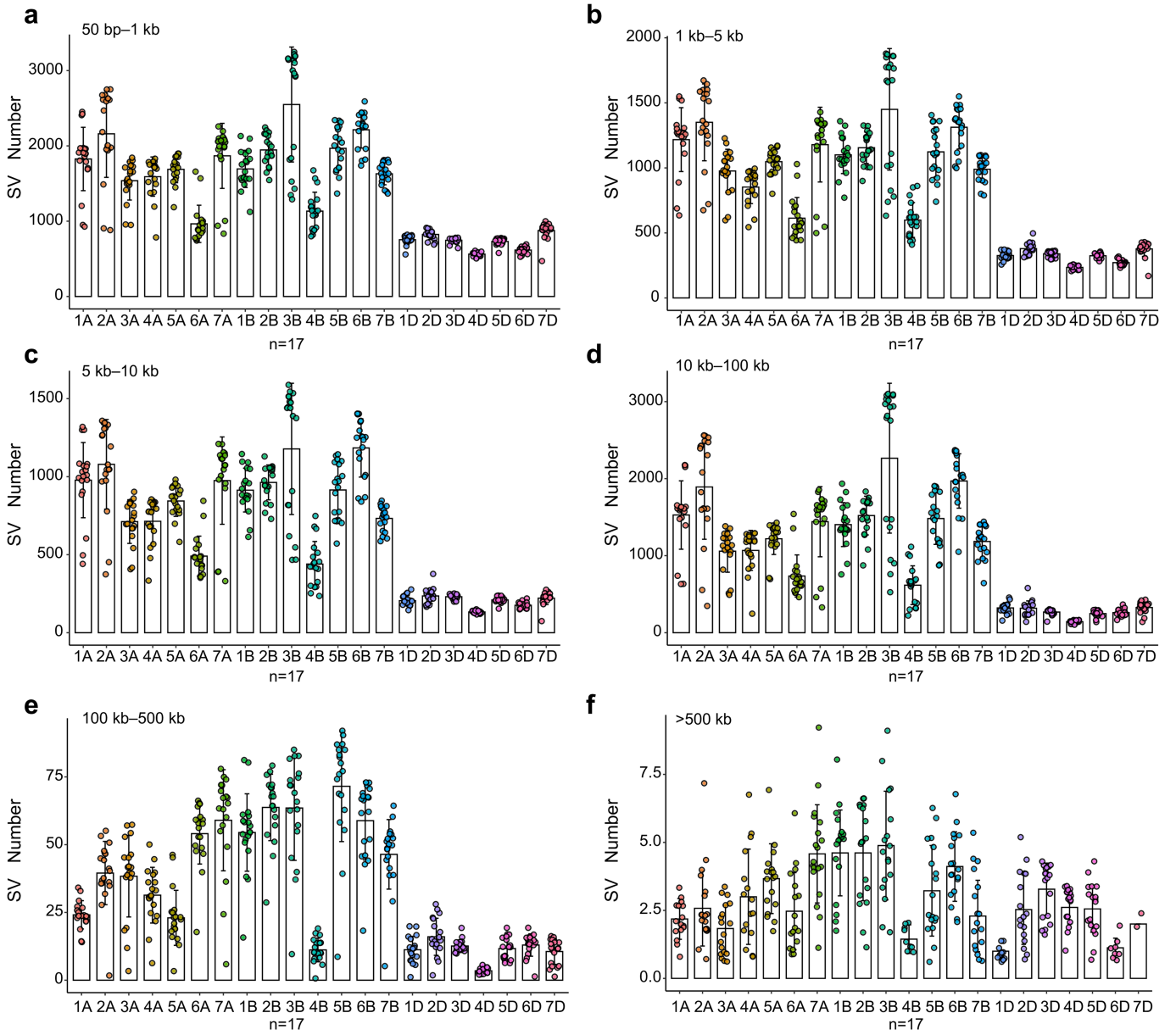
Extended Data Fig. 2 | An overview of the 21 wheat genomes. Living habit of the 20 cultivars with de novo genomes assembly. From left to right, the features are: wheat's spring-winter characteristics, contig N50 and total length of genome, LAI value, and BUSCO evaluation results.



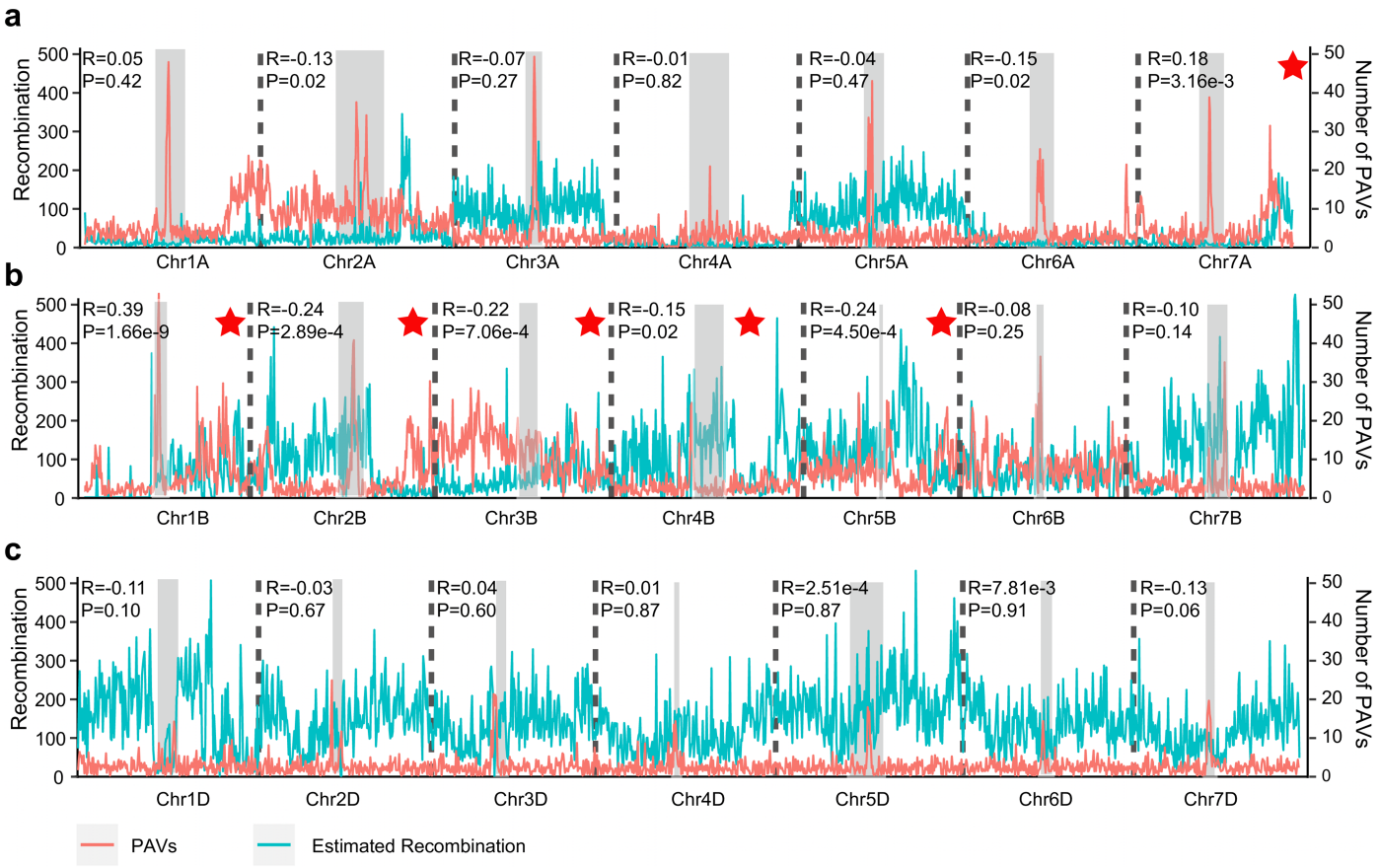
Extended Data Fig. 3 | Wheat pangenome of the 21 cultivars and structural variations referenced to CS. a, Core and pan gene clusters of 20 wheat genomes. The UpSet plot illustrates the core gene clusters (present in all genomes), soft-core gene clusters (present in 18-20 genomes) and dispensable gene clusters (present in 2-17 genomes). **b**, Pie Chart shows the proportion of the gene families marked by each composition. **c**, Density distribution of PAVs in representative cultivars from three breeding ages across chromosomes 3A, 3B, 3D, 5A, 5B, and 5D. The horizontal axis of each box represents the chromosomal position, while the vertical axis indicates the density of PAVs within 1 Mb windows. **d**, Scatter plots showing PAVs occurrence frequencies in 50s&60s and 80s&90s, 80s&90s and post 00s cultivars, respectively. The red scatter indicates the significantly selected PAVs harbored adjusted *P* values (FDR) bigger than 0.01, the green scatter indicates *P* values smaller than 0.01 and bigger than 1e−8, the bule scatter indicate *P* values smaller than 1e−8. The PAVs with frequency more than 0.5 in 80s&90s but less than 0.5 in 50s&60s are defined as specific high-frequency PAVs in 80s&90s. **e**, Source of specific high-frequency PAVs in cultivars released at three breeding stages respectively. The main source of PAV with specific high frequency was European cultivars.



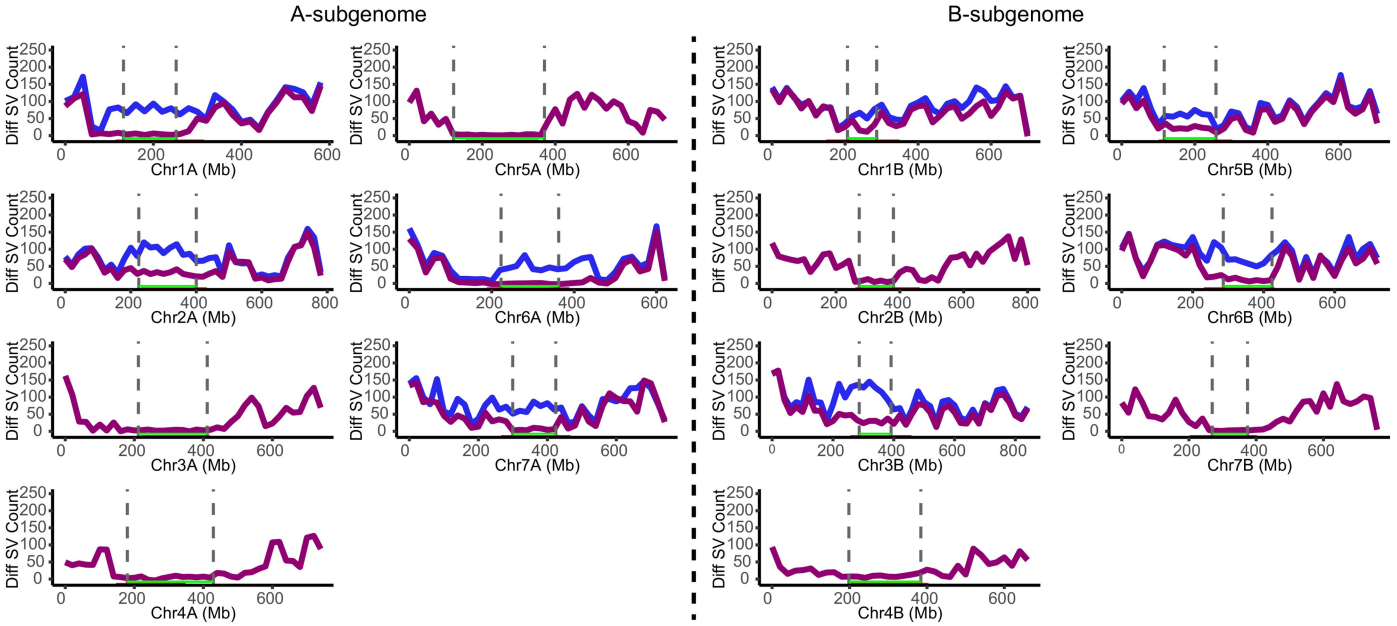
Extended Data Fig. 4 | Features of structural variations (SVs) in the pangenome comprising of 17 wheat cultivars. a, Pie chart shows the proportion of the different SV types in wheat pan-genome. **b**, Percentage of PAVs overlapping with different genomic features. **c**, Distribution of PAVs based on their allele frequency in wheat accessions indicated most PAVs were present in one or only a few accessions. **d**, Distribution of PAV hotspot regions on 21 chromosomes. The red color represents the top 10 regions with PAVs hotspots, and the orange color represents the the next 10-20 regions with a high density of PAV hotspots. **e**, The number of PAVs sharply decreases with PAV length indicated that longer PAVs are relatively rare in the genome, while shorter PAVs are more common. **f**, Characteristic distribution of structural variation lengths in the 17 wheat genomes.



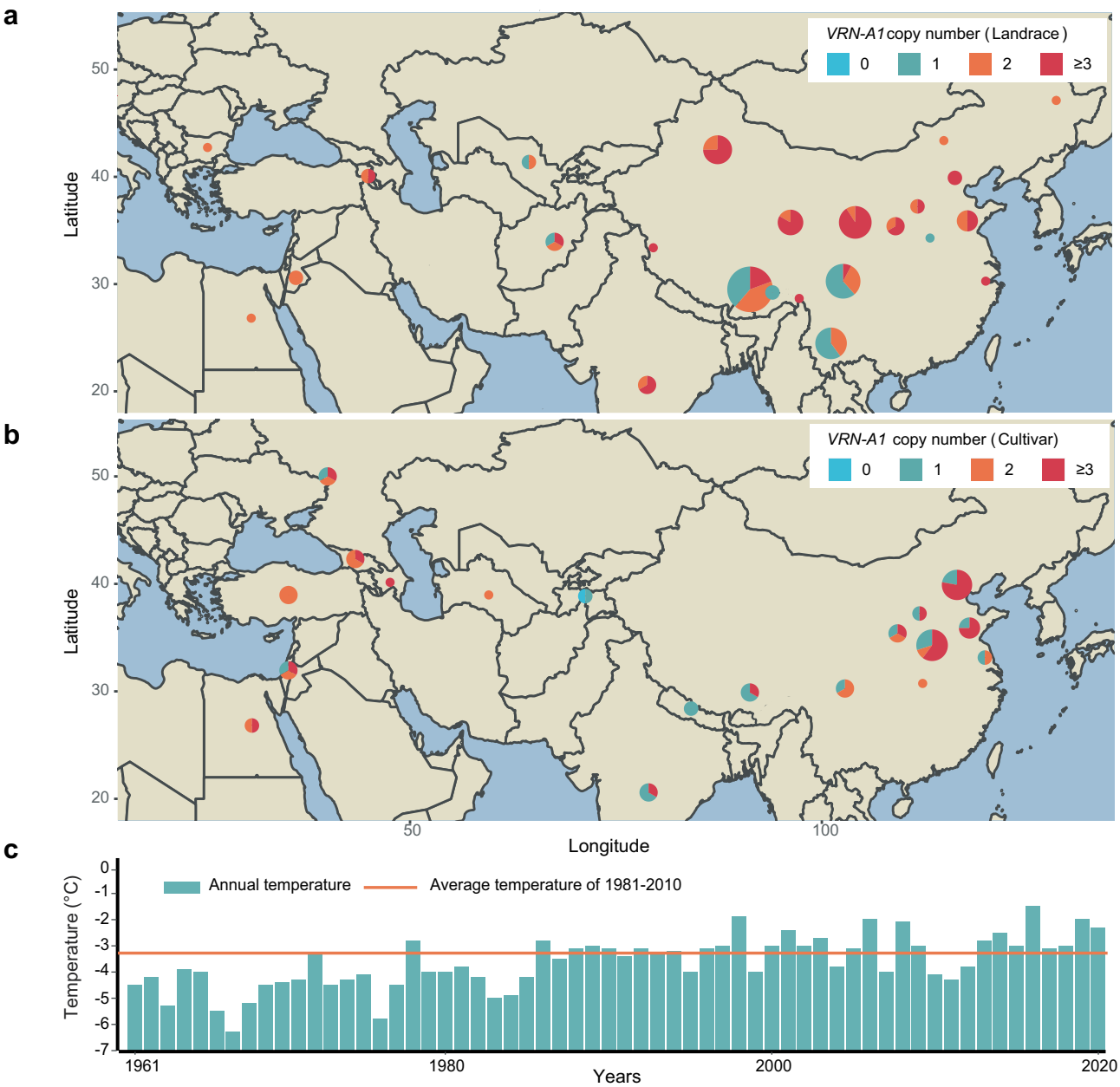
Extended Data Fig. 5 | Distribution of SVs on each chromosome under different length intervals. The B subgenome shows significantly higher density of SVs in the 50 bp–500 kb range compared to the A and D subgenomes, except for the 6A and 4B regions. The D subgenome has a lower density of SVs in this range. No significant difference in the number of SVs was observed among chromosomes for length ranges above 500 kb. **a**, 50 bp–1 kb; **b**, 1 kb–5 kb; **c**, 5 kb–10 kb; **d**, 10 kb–100 kb; **e**, 100 kb–500 kb; **f**, >500 kb.



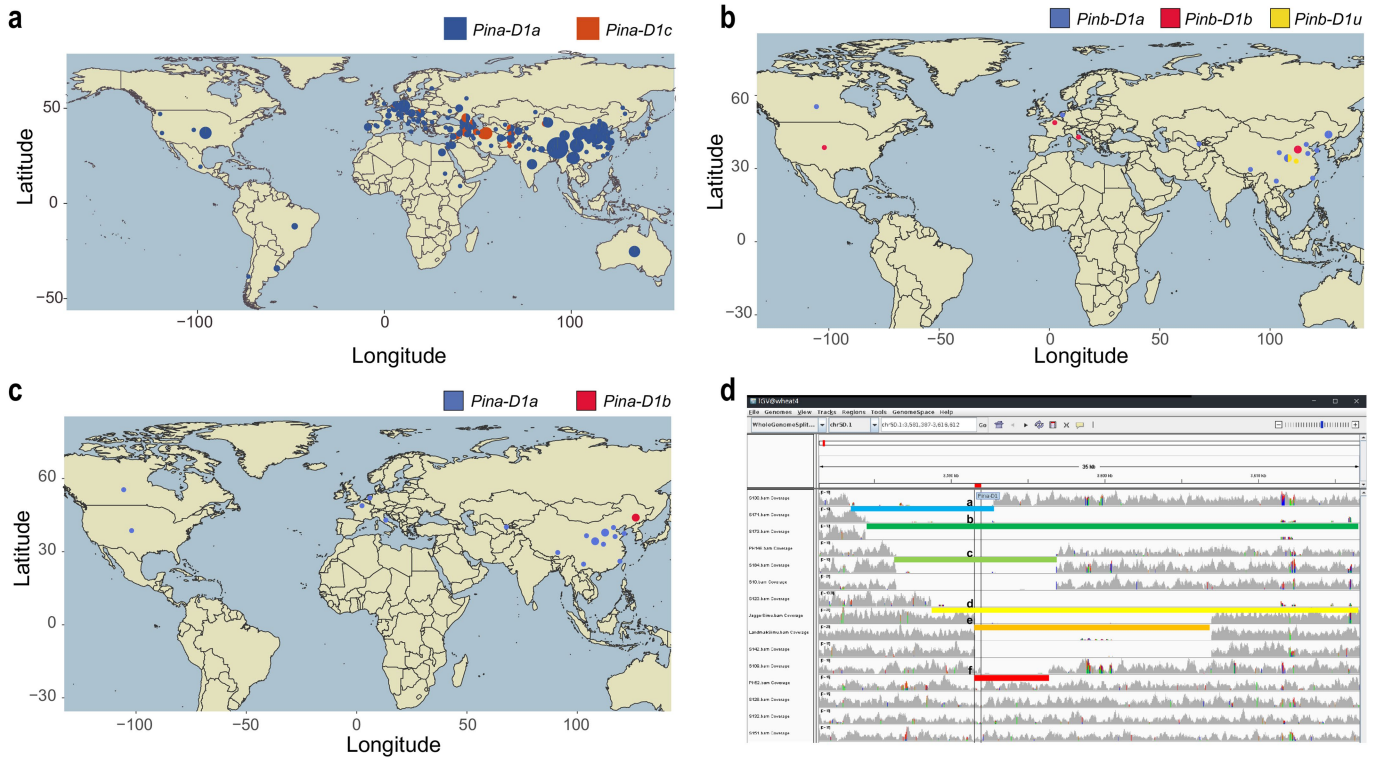
Extended Data Fig. 6 | Crossover recombination at regions proximal to cenctromeres was negatively affected by PAVs strongly. a, **b** and **c** respectively depict the breeding history crossover recombination number (CRN) and PAV number within 100 Mb up- and down-stream of centromeres respectively in the A, B and D sub-genomes. The grey-shaded regions represent the centromeres regions identified by peaks of *CRWs* and *Quintas*. The blue line represents CRN per Mb window estimated based on re-sequencing data of the 145 cultivars in China, while the red line indicates the number PAVs per Mb window in the de novo assembled 17 genomes, representing the wheat breeding history in China. R, Pearson's correlation coefficient between CRN frequencies and PAV counts. Red stars represent significant negative correlation between CRNs and the numbers of PAVs ($P < 0.05$).



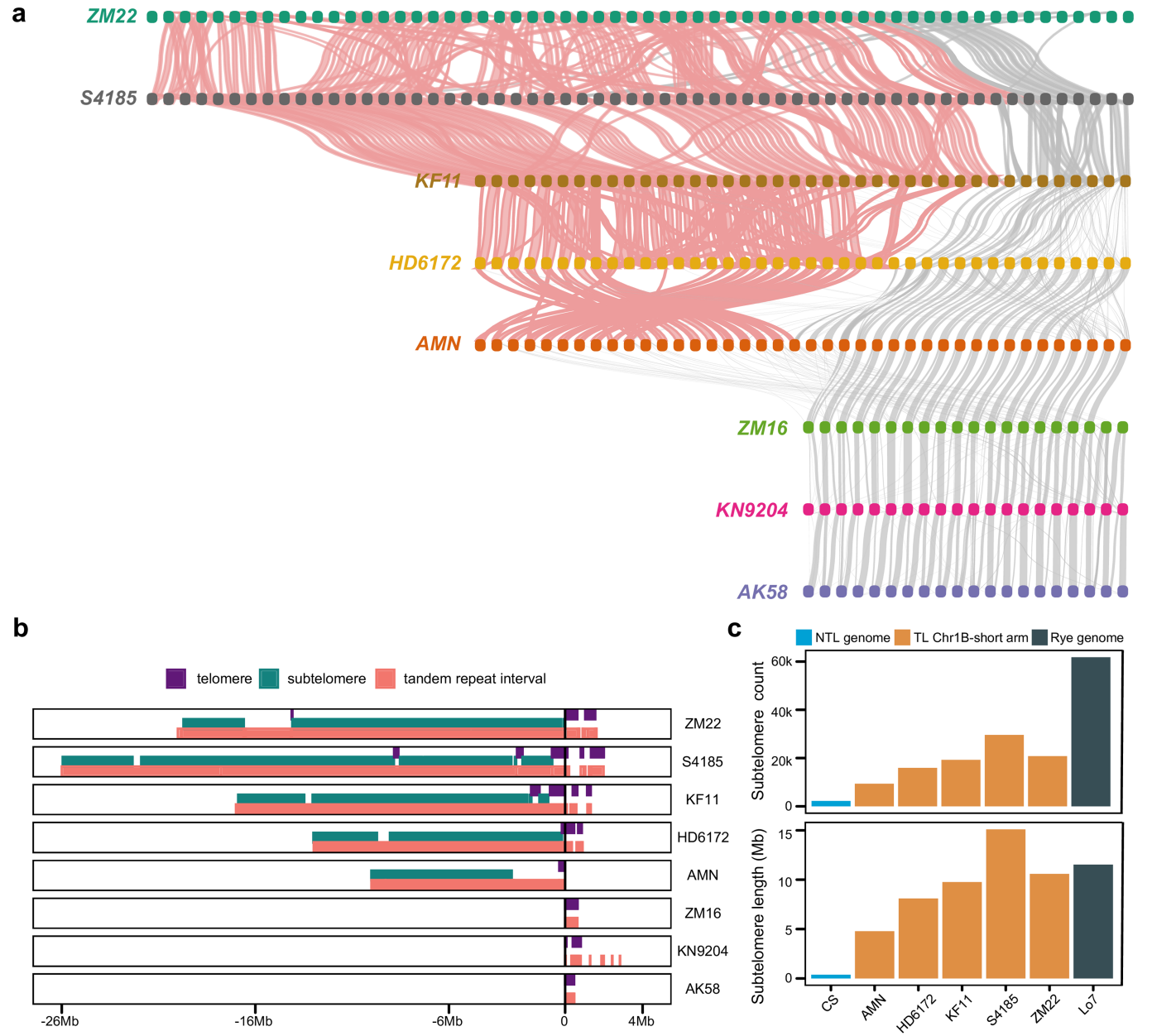
Extended Data Fig. 7 | Partition of and distribution of SVs at regions proximal to centromeres on chromosomes in A- and B-subgenomes. Counts of different SVs between assembly pairs for intra- (purple) and inter-centAHG group (blue) on each chromosome for A- and B-subgenomes. Red bar, regions-proximal to centromeres, from 100 Mb upstream to 100 Mb downstream of centromere on each chromosome. Green bar, centAHG block previously identified[32].



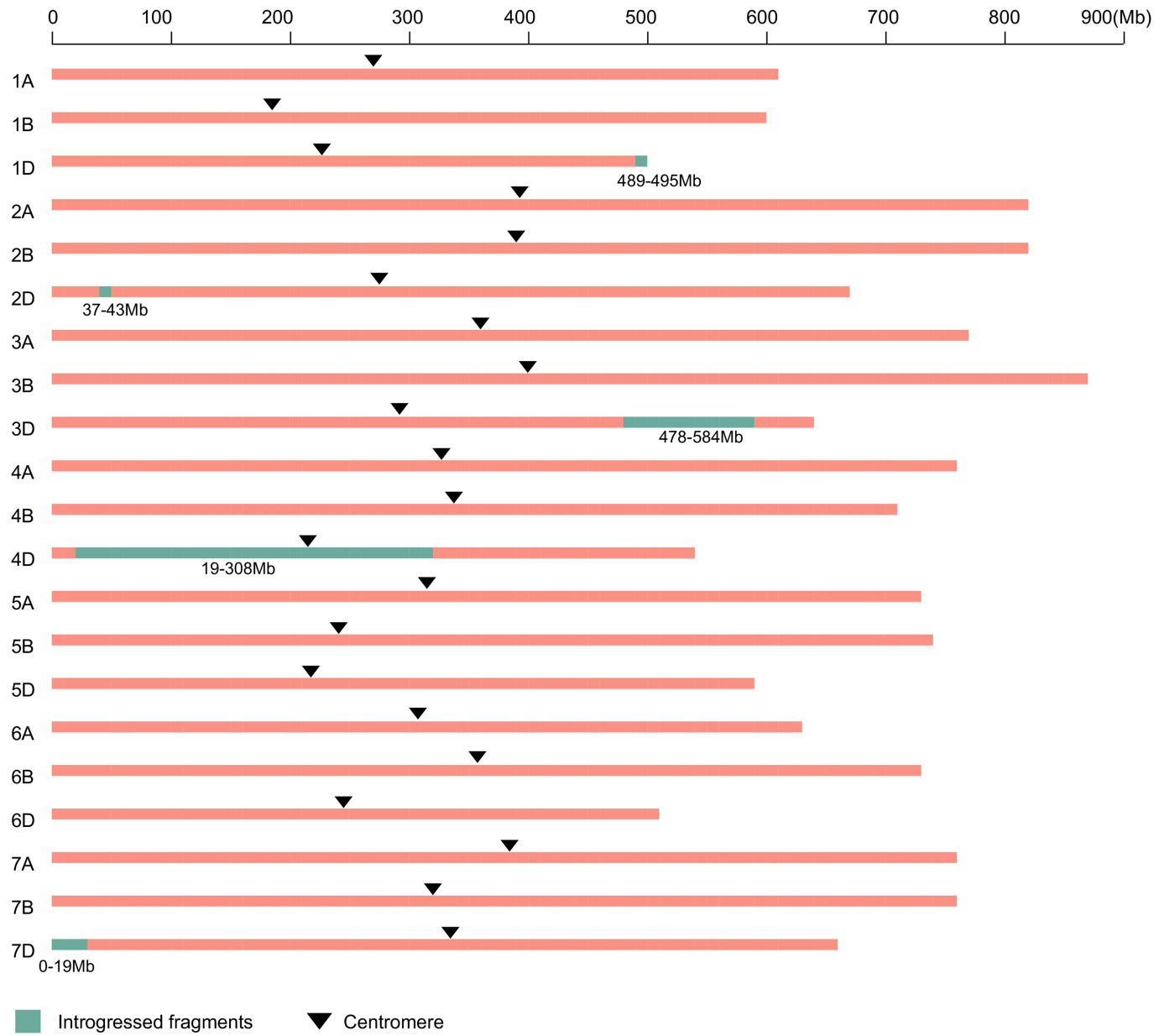
Extended Data Fig. 8 | CNVs in the re-sequenced database of wheat and its tetraploid ancestors indicated *VRN-A1* gene experienced strong selection in wheat origin, spread and breeding in reaction to the environments. **a-b**, Hapmap of *VRN-A1* copy number in landrace and cultivar. Blue represents *VRN-A1* gene loss (0 copy), green, orange, and red represent *VRN-A1* being 1 to 3 copies respectively. During the spread of wheat to China, the copy number increased (triple copy proportion increased), and the share of triple copy types of *VRN-A1* significantly increased in the colder north-western region. However, in modern cultivars the frequency of cultivars with one or two copies increased in north China, probably caused by temperature global warming. The sizes of pies are relative to the count of samples in each location. **c**, Mean temperature in January from 1961 to 2020 in Henan, the largest wheat production province in China. The average temperature in January from 1981 to 2020 is marked with an orange line.



Extended Data Fig. 9 | Gene structure of *Pina* mutations and the geographic distribution of *Pina* and *Pinb*. **a**, The *Pina-D1c* allele found in the resequencing data exists only in Central Asia region. **b-c**, The *Pina-D1b* and *Pinb-D1u* alleles spread mainly to the east. **d**, IGV plot of resequencing data of wheat varieties. The deletion of the fragment results in the disruption of the *Pina* gene structure, followed by gene deletion.



Extended Data Fig. 10 | Large PAVs of sub-telomere repeat on 1RS translocations among modern cultivars. a, The collinearity at the end of short arm of 1RS in cultivars with 1RS·1BL and 1RS·7DL translocation. Each block represents 500 kb of sequence, and to highlight the position of the extended sequences, the collinear relationship between the extended sequences is in light-red. **b**, Distribution of different elements in extended intervals. Purple, green and light red rectangles denoting the telomere-associated sequences (TASs), sub-telomeric location sequences and tandem repeat interval, and the positions are staggered up and down to distinguish different elements intervals. **c**, The number and length of sub-telomere sequences on each genome. The yellow bars represent the number and length of sub-telomere sequences at the end of the short arm of chromosome 1B in 1RS·1BL translocation. The blue and black bars represent the number and length of sequences in the entire genome of non-1RS·1BL translocation hexaploid wheat and rye, respectively.



Extended Data Fig. 11 | Genome composition of CM42, the first cultivar derived from cross between common wheat and the CIMMYT synthetic. Introgressed fragments from *Aegilops tauschii* are marked in green. Cetromeres were indicated by black triangles.

Extended Data Table 1 | Statistics of the assembly and annotation of 21 wheat genomes

Accession	Assembly length	Contig N50	Number of contig in chromosome	Gap number	Chromosome ahchoring rate(%)	Busco	LAI	Corrected Gene number	Breeding age	Ecotype
New assembly										
BJ8	14,946,041,461	19,242,618	2,306	2,285	97.08%	99.10%	16.40	154,055	50-60s	Strong Winter
MZM	14,672,384,725	32,613,137	1,620	1,599	98.56%	98.90%	16.59	152,776	50-60s	Strong Winter
XN6028	14,692,619,227	31,752,513	1,325	1,304	98.64%	99.00%	16.40	153,509	50-60s	Strong Winter
Abo	14,702,933,806	25,880,799	1,357	1,336	97.97%	99.00%	16.38	151,829	80-90s	Semi Winter
NC4	14,576,931,732	31,333,180	1,367	1,346	98.79%	98.90%	16.36	151,701	80-90s	Spring
YM158	14,702,635,128	25,798,185	2,016	1,995	98.38%	99.00%	16.46	153,162	80-90s	Semi Winter
XY6	14,873,647,137	25,461,158	1,328	1,307	97.11%	99.00%	17.20	153,075	80-90s	Winter
AMN	15,006,113,569	24,549,188	1,394	1,373	96.82%	99.00%	16.51	152,478	80-90s	Strong Winter
JM47	15,045,816,769	21,604,926	5,406	5,385	98.27%	99.00%	16.41	154,828	80-90s	Strong Winter
S4185	14,992,435,260	26,765,449	2,811	2,790	97.15%	99.00%	16.51	152,811	80-90s	Strong Winter
CM42	14,656,108,654	41,616,920	988	967	98.18%	98.90%	16.36	151,806	post-00s	Semi Winter
JM22	15,041,748,690	26,063,057	1,170	1,149	95.91%	99.00%	16.40	152,462	post-00s	Winter
KF11	14,855,443,168	30,316,820	1,282	1,261	97.58%	99.10%	16.54	151,964	post-00s	Winter
ZM366	14,829,756,710	26,279,961	3,933	3,912	97.78%	99.00%	16.77	153,198	post-00s	Winter
ZM16	15,063,560,237	33,180,015	3,068	3,047	96.62%	99.00%	16.55	154,524	post-00s	Winter
ZM22	15,041,748,690	26,063,057	2,872	2,851	96.76%	99.00%	16.66	153,111	post-00s	Winter
HD6172	15,099,169,769	26,796,172	2,572	2,551	96.23%	99.00%	16.67	154,470	post-00s	Strong Winter
Published assembly										
CS	14,577,412,364	341,062	306,746	306,724	97.59%	98.70%	14.97	152,169		Spring
Fielder	14,702,880,414	20,684,258	1,428	1,407	97.86%	98.70%	14.30	151,846		Spring
Kariega	14,677,204,660	26,664,824	2,487	2,466	98.47%	99.20%	16.87	152,838		Spring
Attraktion	14,244,925,392	17,258,578	1,553	1,532	--	99.10%	16.14	148,999		Winter

nature portfolio

Corresponding author(s): Xueyong Zhang,Weilong Guo and Rajeev K. Varshney

Last updated by author(s): January 4,2024

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	DNA sequencing Pacbio reads were collected from single-molecule real-time SMRT cells on Pacbio Sequel II platform HiC data and paired-end reads were collected from the Illumina NovaSeq 6000 platform. RNA-seq data were generated using Ilumina NovaSeq 6000 platform.
Data analysis	Hifiasm 0.13-r308 HiCUP v0.7.3-1 ALLHiC v0.9.8 Juicebox v1.11.08 NUCmer v 3.1 BUSCO v4.1.4 LTR_retriever v2.6 RepeatMasker v 4.0.5 TRF v409 SOLAR v0.9.6 GeneWise v2.4.1 PASA v2.2.0 Augustus v2.5.5 Genscan v1.0 Geneid v1.4.4 GlimmerHMM v 3.0.1 TopHat v2.0.8

Cufflinks v2.1.1
EvidenceModeler
JCVI
RGAugury
NLR-annotator v.0.7
BWA v0.7.17-r1188
SAMtools v0.1.1
Sentieon
GATK
ANNOVAR v 2013-05-20
SyRI
SURVIVOR
svtyper
VCFtools v0.1.14
EMMAX package
MAFFT v7.471
FastTree v 2.1.11
bedtools v2.26.0
Mclust v 6.0.1
GeneTribe
GenomeSyn v1.41
RepeatMasker v4.0.7
minimap2 v2.22-r1110

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

Provide your data availability statement here.

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	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used.* *Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected.* *Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Reporting on race, ethnicity, or other socially relevant groupings	*Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status).* *Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.)* *Please provide details about how you controlled for confounding variables in your analyses.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

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

Life sciences study design

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

Sample size	17 representative accessions collected from China were selected for denovo genome assembly. The 145 accessions from the historical series of Chinese wheat breeding included 100 modern Chinese cultivars (MCC), 25 Chinese landraces (CL), and 20 elite cultivars were selected for population analysis.
Data exclusions	No data were excluded.
Replication	All experiments were repeated at least two or three times, and the number of independent experiments or biological replicates is indicated in the figure legends
Randomization	All samples were collected randomly into experimental groups. The plant materials were grown under specific conditions and planting methods, which are described in detail in the methods.
Blinding	The research materials are plants so the blind design is not applicable in the field. For molecular biology experiments, bias could not be introduced since samples were treated identically and collected randomly. Experiments were repeated by different authors. The researchers also evaluated agronomic traits and performed RNA-seq analysis without prior knowledge of the results.

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

Plants

Seed stocks	Fifteen of the 17 wheat cultivars de novel sequenced and assembled were kept in China National Gene Bank with following ids, which can be obtained after signature of MTA following Chinese Government Legislation of Crop Germplasm (www.cgris.net).
Novel plant genotypes	Beijing 8 (ZM8963) , Mazhamai (ZM3697), Xinong 6028 (ZM9597), Abo (MY161), Ningchun 4 (ZM017424), Yangmai 158 (ZM025136), Xiaoyan 6 (ZM017079), Aimenniu (ZM015830), Jinmai 47 (ZM030317 Shimai 4185 (ZM025983), Chuanmai 42 (ZM027007), Jimai 22 (ZM028387), Zhengmai 366 (ZM026975), Zhoumai 16 (ZM026962), Zhoumai 22 (ZM027068). Two cultivars, Handan 6172 and Kuofan 11 have not been cataloged in the national crop germplasm information system, which can be provided under special agreement following the Chinese Government Legislation of Crop Germplasm.
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*