Journal of Genetics and Genomics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Genetics and Genomics

Journal homepage: www.journals.elsevier.com/journal-of-genetics-and-genomics/

Original Research

Structural variation-based and gene-based pangenome construction reveals untapped diversity of hexaploid wheat

Hong Cheng [a, b, *, 1], Lingpeng Kong [a, 1], Kun Zhu [c, 1], Hang Zhao [a], Xiuli Li [a], Yanwen Zhang [a], Weidong Ning [a], Mei Jiang [a], Bo Song [a], Shifeng Cheng [a, *]

[a] Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong 518100, China
[b] College of Grassland Agriculture, Northwest A&F University, Yangling, Shaanxi 712100, China
[c] State Key Laboratory of Crop Stress Adaptation and Improvement, School of Life Sciences, Henan University, No. 379 Mingli Road (North Section), Zhengzhou, Henan 450046, China

ARTICLE INFO

Article history:
Received 28 November 2024
Received in revised form
25 March 2025
Accepted 27 March 2025
Available online xxx

Keywords:
Wheat
Pangenome
Structural variation
Centromere plasticity
Growth habit

ABSTRACT

Increasing number of structural variations (SVs) have been identified as causative mutations for diverse agronomic traits. However, the systematic exploration of SVs quantity, distribution, and contribution in wheat was lacking. Here, we report high-quality gene-based and SV-based pangenomes comprising 22 hexaploid wheat assemblies showing a wide range of chromosome size, gene number, and TE component, which indicates their representativeness of wheat genetic diversity. Pan-gene analyses uncover 140,261 distinct gene families, of which only 23.2 % are shared in all accessions. Moreover, we build a ~16.15 Gb graph pangenome containing 695,897 bubbles, intersecting 5132 genes and 230,307 cis-regulatory regions. Pairwise genome comparisons identify ~1,978,221 non-redundant SVs and 497 SV hotspots. Notably, the density of bubbles as well as SVs show remarkable aggregation in centromeres, which probably play an important role in chromosome plasticity and stability. As for functional SVs exploration, we identify 2769 SVs with absolute relative frequency differences exceeding 0.7 between spring and winter growth habit groups. Additionally, several reported functional genes in wheat display complex structural graphs, for example, *PPD-A1*, *VRT-A2*, and *TaNAAT2-A*. These findings deepen our understanding of wheat genetic diversity, providing valuable graphical pangenome and variation resources to improve the efficiency of genome-wide association mapping in wheat.

Introduction

Massive efforts have been devoted to assembling the genome of allohexaploid wheat (*Triticum aestivum*; AABBDD) in recent years (Brenchley et al., 2012; IWGSC, 2014; Clavijo et al., 2017; Zimin et al., 2017), because of its incredibly large size and high complexity. The chromosome-scale and fully annotated wheat genome (IWGSC RefSeq v1.0) using Chinese Spring (CS) with high completeness and accuracy has been released in 2018 (IWGSC, 2018). Since then, a growing number of high-quality genomes for different wheat breeds have been published and updated (Guo et al., 2020; Walkowiak et al., 2020; Athiyannan et al., 2022; Aury et al., 2022; Kale et al., 2022; Jia et al., 2023), accompanied by massive accumulation of Next-Generation Sequencing (NGS) data of common wheat and their wild relatives (Cheng et al., 2019, 2024; Pont et al., 2019; Guo et al., 2020; Zhou et al., 2020; Wang et al., 2022), providing essential support for genomic diversity exploration of this species. These researches mostly used a single reference genome, while it has been widely accepted that the single genome could not fully capture the variations within species and further limits the ability to capture population genetic diversity (Bayer et al., 2020). Pangenome, representing the collection of all genetic diversity in a species, rapidly developed and seemed to be a new reference for genome research and breeding (Shi et al., 2023; Schreiber et al., 2024), especially the graph-based pangenome that represents tens and hundreds of alleles simultaneously without bias (Eizenga et al., 2020).

Compared with other crops, the pan-genomic research on wheat is still quite inadequate. The first attempt to construct wheat pangenome using NGS data from 18 wheat cultivars revealed that 64.3%

* Corresponding authors.
E-mail addresses: chenghong92@163.com (H. Cheng), chengshifeng@caas.cn (S. Cheng).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jgg.2025.03.015

Please cite this article as: H. Cheng, L. Kong, K. Zhu *et al.*, Structural variation-based and gene-based pangenome construction reveals untapped diversity of hexaploid wheat, Journal of Genetics and Genomics, https://doi.org/10.1016/j.jgg.2025.03.015

of genes were shared by all cultivars, and only 245 genes were absent in CS reference (Montenegro et al., 2017). The size of wheat pangenome was undoubtedly underestimated in this study owing to the relatively narrow set of cultivars. More recent exploration using 10+ Wheat Genome lines assembled a 15.8 Gb graph and estimated only 19% of segments were present in all accessions (Bayer et al., 2022). In addition, our previous research based on pseudo-genome made up of wild emmer and *Ae. tauschii* genome retrieved 1517 Mb sequence absent in CS (Cheng et al., 2019). Not taking into account the sequence absent in CS would lead to a huge amount of diversity being missed. Therefore, the construction of wheat pangenome and subsequent in-depth investigation of its genetic diversity based on pangenome are extremely required.

Beyond SNPs and Indels, GWAS and QTL have identified numerous structural variations (SVs) as causative variations for important agronomic traits of plant genome, such as fruit flavor, size, and production in tomato (Alonge et al., 2020), fruit spine in cucumber, and heat tolerance in pearl millet (Yan et al., 2023). Graph-based pangenomes efficiently accelerated the identification of major SVs underlying diverse phenotypes (Zhou et al., 2022; Lyu et al., 2023; Yan et al., 2023). Up to now, little is known about wheat SVs. The most prominent SVs identified in wheat were large-scale translocations and inversions between chromosomes (Walkowiak et al., 2020). Although the sequencing technology and bioinformatic method have greatly advanced, it is still challenging to characterize SVs accurately in such an intricate genome. A latest study provided a refined SV set including 1,030,303 deletions and 1,457,547 insertions in wheat (Zhang et al., 2024) and thoroughly evaluated the approaches for SV identification. Taken together with the accumulation of high-quality wheat genome collections, constructing graph pangenome and interpreting patterns of genetic diversity using SVs became achievable in wheat. However, a high-quality graph pangenome that can be widely used in wheat genomic research is lacking.

In this study, we constructed a gene-based and a graph-based pangenome by adequately combining 22 assemblies from a diverse range of wheat lines (Table S1), in contrast with the more limited scope in previous studies (Montenegro et al., 2017; Jiao et al., 2024). We further detected millions of SVs, including PAVs, translocations, and inversions, and explored the distribution of SV hotspots across the whole genome. To seek out functional SVs underlying such a large dataset, we then calculated the frequency difference between Spring and Winter wheat populations. In addition, we showcased and resolved the subgraph of reported functional genes contributing to phenotypic variations in wheat (Rasheed et al., 2024). Our graph-based pangenomes are freely available at http://wheatpgdb.cn/, where complex local graphs are better presented compared to the linear pangenome. WheatPGDB will provide an important public resource for wheat diversity exploration and genomics research. Our attempt to perform graph pangenome construction and SV identification in such a large, complex, and polyploid genome also provides an example for genomic analysis in other comparable species.

Results

Evaluation and statistics of multiple wheat assemblies

We first collected 17 chromosome-level and 5 scaffold-level of well-known hexaploid wheat assemblies from previous publications (IWGSC, 2018; Guo et al., 2020; Walkowiak et al., 2020; Sato et al., 2021; Athiyannan et al., 2022), covering several top 10 wheat-producing countries worldwide (FAO, 2022), including China, America, France, Canada, Germany, and Australia (Fig. 1A; Table S1). Their phylogenetic relationship was mainly driven by the geographic distribution (Fig. 1B), consistent with the patterns reported previously (Cavanagh et al., 2013; Cheng et al., 2019). The assembly sizes of these genomes range from 13.9 Gb to 15.0 Gb (including chromosome Un or unanchored contigs) (Fig. 1B), with high completeness supported by a mean of 99.1% (98.5%−99.3%) Benchmarking Universal Single-Copy Ortholog (BUSCO) (Simão et al., 2015) (Fig. S1). The transposable elements (TEs) made up 77.4%−86.8% of these genomes, of which, on average, 66.4% were long terminal repeat (LTR) and 5.8% were terminal inverted repeat (TIR) (Figs. 1B and S2; Table S2). The contig N50 and scaffold N50 showed a wide range because of the difference in resequencing technology; contig N50 were 16.4 kb−30.2 Mb, and scaffold N50 were 10.2 Mb−717.2 Mb (Fig. 1C). In consideration of the assembly quality and potential practicability in the breeding process integrally, we chose Fielder, an American spring cultivar known for its amenability to genetic editing (Sato et al., 2021), as the backbone of wheat graph pangenome.

Among these assemblies, the chromosome-level genome assemblies exhibited a wide range of chromosome sizes (Fig. 1D). The length of chromosome 5B can differ by as much as 258.5 Mb (Table S3), ranging from 467.8 Mb to 726.4 Mb. We performed alignment genome by genome to explore the contributing factors to such large-scale variation in chromosome size. We observed that the most prominent event was translocations between different chromosomes and pericentric inversions within chromosomes (Figs. 1E, S3−S8). Among them, 5B/7B translocation (Figs. 1E and S3) in SYMattis and ArinaLrFor, 2B translocation in Lancer, inversion on 4B (Fig. 1E) in Attraktion, Lancer, Landmark, Renan and Spelt, and inversion on 5B in Julius have been researched previously (Walkowiak et al., 2020). Despite these, we still found several unexplored large-scale structural variations between pairwise alignments, including translocations between 7B and 7D, 1B and 1D, 6D and 6B, and inversions on 3D and 7B (Figs. S3−S8). These large-scale interchromosomal and intrachromosomal rearrangements, together with observed variations in genome size and component (such as gene and TE), highly indicated the representativeness and unexplored genic diversity of these 22 genomes to build a wheat graph pangenome.

A protein-coding gene-based pangenome

Pan-gene map, which clustered all available genic sequences, provided a complete picture of the presence-absence variations (PAVs) of genes within a species. Gene PAV was complementary to other genetic markers in mapping functional genes. Variable genes have been frequently predicted to be associated with adaptation and agronomic traits (Li et al., 2014; Golicz et al., 2016; Hubner et al., 2019). Therefore, we performed an exploration of the PAVs of protein-coding genes prior to the construction of SV-based graph pangenome (see Materials and methods). Three accessions (Attraktion, Kariega, and Renan) were excluded, one for the lack of genome annotation and the other two for low annotation quality measuring by BUSCO (52.4% and 5.8%) (Fig. S1). A total of 106,914 to 145,065 protein-coding genes have been annotated in each of the 19 remaining assemblies previously (Table S1). The number of added gene families declined rapidly with the genome number increased, and the curve of the pan-gene family number nearly reached a plateau when the number of accessions was greater than or equal to 17 (Fig. 2A). Only ~310 additional clusters were detected when adding the 18th accession, indicating that our set of 19 accessions legitimately captured most of the gene family diversity in wheat.

Based on the frequency of gene families across 19 genomes, we classified gene families into four categories, as described in a previous study (Kang et al., 2023). Finally, we constructed a gene-based pangenome formed by 140,261 non-redundant pan-gene clusters of

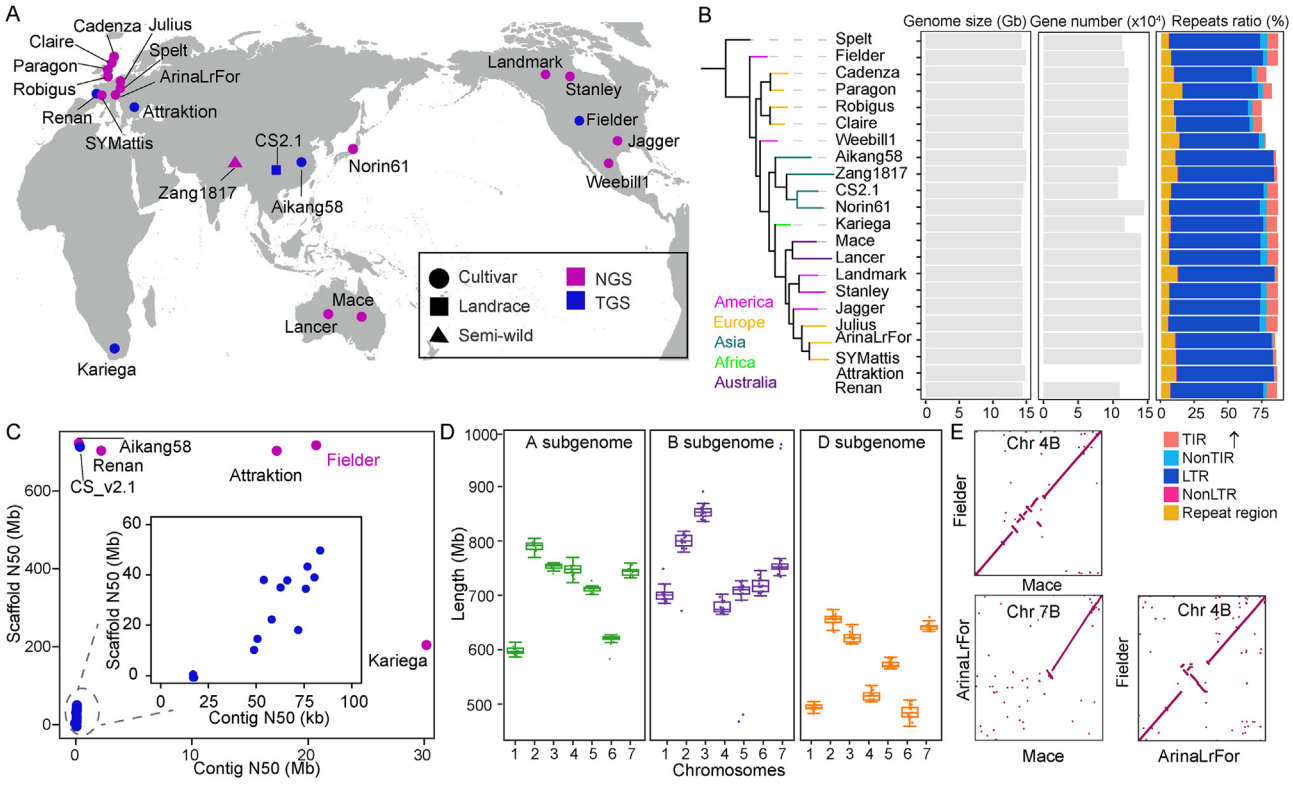

Fig. 1. Overview of geographic distribution, genome assembly statistics, and comparison of 22 wheat assemblies. **A**: Worldwide distribution of the 22 wheat assemblies. The color and shape of the points represent the sequencing platforms and improvement status, respectively. The map is accessible at http://www.tianditu.gov.cn with a map ID GS(2016)1611. **B**: Phylogenetic relationship followed by genome size, gene number, and TE component. Phylogenic tree was built using sequences of 6855 orthogroups identified by OrthoFinder. There was no available annotation for Attraktion. **C**: The assembly statistics, including contig and scaffold N50 of 22 assemblies. **D**: The variations of chromosome length among different assemblies. **E**: Large-scale interchromosomal and intrachromosomal structural variations revealed by pairwise genome alignments.

protein-coding genes, of which 32,529 (23.19%) were core genes that present in 19 genomes, 19,654 (14.01%) were softcore that shared by 90 % genome (n = 17 to18) but not all, 83,773 (59.73%) were dispensable that shared by 1−16 genome, and 4305 (3.07%) were private that present in only one genome (Fig. 2B). The core gene percentage was quite low as compared with that reported in previous studies using genomes with limited scope of geographic origin, 64.3% for 18 cultivars mainly from Australia (16/18) (Montenegro et al., 2017), 65.66% for 17 cultivars mainly from China (16/17) (Jiao et al., 2024). We noticed that the frequency distribution of gene families in wheat was quite different from that in other species, which pictured a U-shaped distribution in general (Collins and Higgs, 2012; Liu et al., 2020; Chen et al., 2023; Shi et al., 2024; Zhang et al., 2024). On average, core genes occupied 40.76% (36.04%−54.60%) of the total genes per accession (Fig. 2C and 2D; Table S4), much lower than that in other species, for example, more than 50% in broomcorn millet (Chen et al., 2023), 80 % in cucumber (Li et al., 2022). In addition, 71,701 (51.12%) gene families in the graph were absent in the CS reference genome.

Construction of the SV-integrated graph

To obtain the additional sequences not contained in the present wheat reference (IWGSC RefSeq v2.1) (Zhu et al., 2021), we used the Fielder linear reference as the backbone of the graph as described above and integrated the other 21 assemblies to construct the graph using minigraph (Li et al., 2020) (see Materials and methods). The resulting pangenome length was ~16.15 Gb, ~1.76 Gb larger than the CS2.1 reference (Zhu et al., 2021), ~0.35 Gb larger than the pangenome based on 16 wheat assemblies (Bayer et al., 2022), confirming the ability to capture additional variations of our graph. To further validate the segments, we aligned the assemblies to this graph, and those supported by at least one assembly were retained. The final graph was split into 2,580,200 segments with an average size of 6.25 kb ranging from 1 bp to 5.29 Mb (Fig. S9; Table S5). Among them, 9.14%, 9.95%, 66.87%, and 14.13% were core, softcore, dispensable, and private, respectively (Fig. S10). The frequency distribution of segments among 22 genomes (Fig. S10) was more like U-shape than gene families (Fig. 2B). We further scanned segments in each genome. The segment numbers of five scaffold-level assemblies were significantly lower than chromosome-level assemblies (Fig. S10), principally resulting from the genome incompleteness induced while establishing pseudochromosomes. However, Weebill1, one of the five scaffold-level assemblies, possessed the largest number of private segments (50,175 segments with a total length of 14.49 Mb). It should be noted that 1,171,211 (45.39%) segments cannot be fully detected in CS2.1 reference; that is, depth of coverage was lower than 0.9 measured by assembly-to-graph mapping (see Materials and methods).

Next, we explored the flexible regions that augment the graph. We identified 695,879 bubbles with a total length of 1536 Mb across 21 chromosomes (Fig. 3A). There was an obvious asymmetric distribution of bubbles among subgenomes with B > A ≫ D (Fig. 3A), consistent with the diversity patterns demonstrated by SNPs, Indels, and CNVs in previous studies (Cheng et al., 2019; Pont et al., 2019). We then decomposed bubbles into their constituent structural haplotypes and investigated different coordination of these haplotypes (alleles/paths). Each bubble contained on average 4.76 segments, with an extreme case in which 723 haplotypes were nested. Among these bubbles, 80.36% (559,226) were biallelic, 13.79% had three to

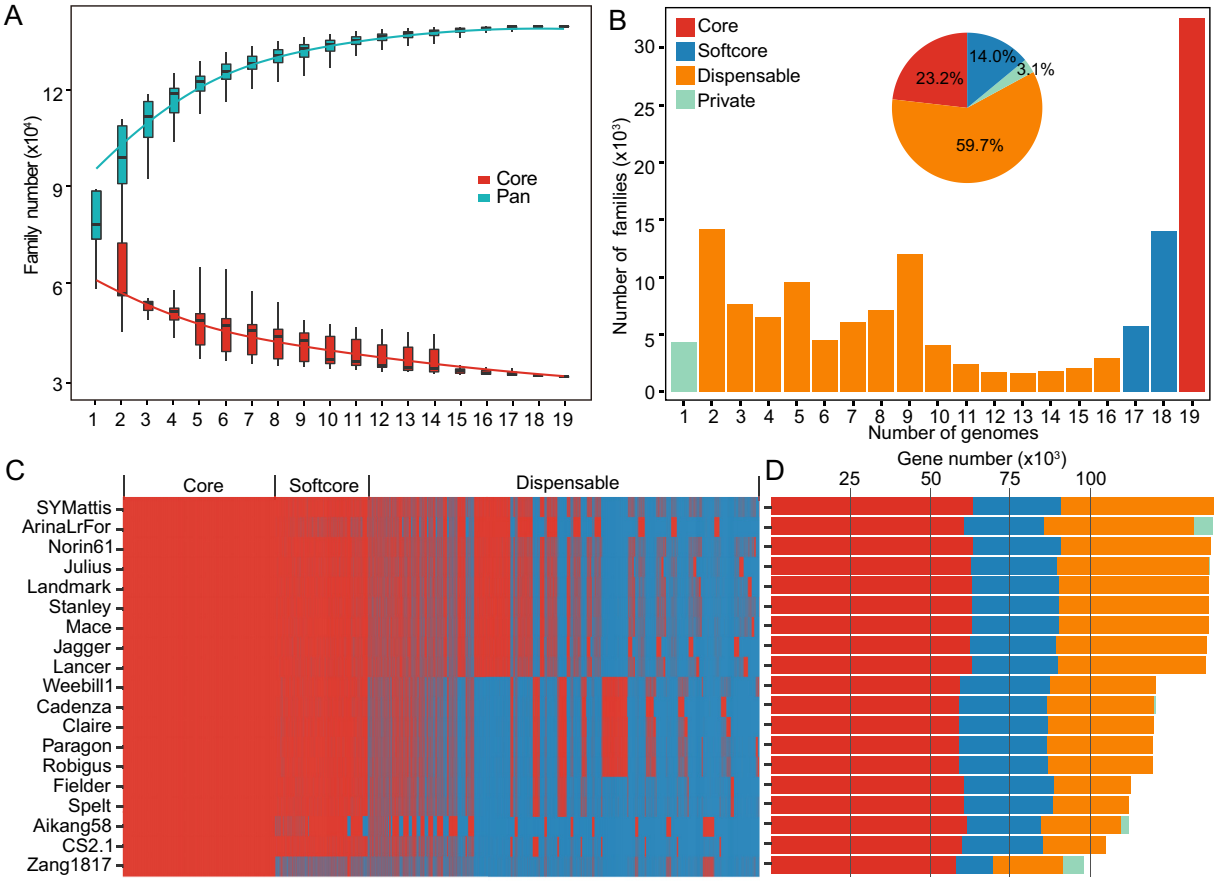

Fig. 2. Summary of gene-based pangenome using 19 annotated genomes. **A**: Simulations of the gene-based pangenome size for different combinations of genomes. Numbers of pan- and core-gene families are displayed in blue and red, respectively. **B**: Distribution of gene families across 19 genomes and composition of core, softcore, dispensable, and private genes. **C**: Landscape of gene PAVs for 19 wheat accessions. **D**: Gene number of core, softcore, dispensable, and private genes of individual genomes.

five alternative alleles, and 5.85% had more than five alleles (Fig. 3B). The percentage of the biallelic bubble was similar to that of genes clustered into no more than three haplotypes (83%) in global wheat (Scott et al., 2021), conformably hinting at limited haplotype diversity. However, the sufficiently diverged subsequences making up the multiallelic bubbles potentially have great value in enriching genetic and phenotypic diversity.

There were on average ~48 bubbles per 1 Mb across the whole genome (Figs. 3C and S11). Of note, we observed that bubble breakpoints display significant enrichment in centromeres (CENs), positioned in IWGSC RefSeq v1.0 (IWGSC, 2018), on almost all the 21 chromosomes (Fig. 3D and Fig. S11). On the contrary, SNPs, Indels, gene models, and their derived nucleotide diversity (π) were consistently low across centromere regions (IWGSC, 2018; Pont et al., 2019; Jia et al., 2023). Functional investigation revealed that a large part of the bubbles were located in the intergenic region (Fig. 3E). Moreover, 8843 bubble breakpoints were located in protein-coding region of 5132 genes in Fielder coordinates (Fig. 3E), of which 4558 were synonymous variations, 380 were stop-gain mutations, and 19 were stop-loss mutations, indicating their apparent effect on variability in the length of protein sequences. The involved genes were annotated to be associated with several important processes, such as phenylpropanoid biosynthesis, which was proved to be associated with insect resistance traits (Wang et al., 2024); ABC transporters, responsible for transmembrane transport of many primary metabolites (Fig. S12). To further explore the potential regulatory impact of these bubbles, we integrated the available ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) data of wheat (Lu et al., 2020; Pei et al., 2023a, 2023b; Zhao et al., 2023b). Using these data, we totally identified 230,307 ATAC-seq peaks. We then overlapped bubble breakpoints with these peaks and observed 32,546 bubble breakpoints were located in 22,435 (9.74 %) ATAC-seq peaks (Fig. 3F). The percentage of core segments was much higher for those overlapped with ATAC peaks than genome-wide (Fig. 3G), hinting at the evolutionary conservation of regulatory sequences.

A comprehensive landscape of SVs

Briefly, each assembly was aligned to the Fielder genome, and then poorly mapped sequences with identity <90% identity and with lengths <200 bp were excluded. SVs were identified in all accessions using SyRI (Goel et al., 2019) (Figs. S13–S15). On average, each accession has 368,430 (100,929–596,971) SVs with a length exceeding 50 bp (Fig. S16; Table S6) and then merged by subclassification to construct a comprehensive SVs map of wheat and to enable exploration at the population level. Consequently, we obtained a set of 1,978,221 non-redundant SVs (>50 bp) (Figs. S17 and S18) and 334,035 NOTAL (Un-aligned region) variations that cannot be merged because of inconsonant query coordinates. Of the total SVs set, 60.3% were detected in only one accession (Fig. 4A). We further examined each SV type across six main length gradients: 50 bp–100 bp, 100 bp–500 bp, 500 bp–1 kb, 1 kb–10 kb, 10 kb–50 kb, and >50 kb. The percentages for length gradients were largely

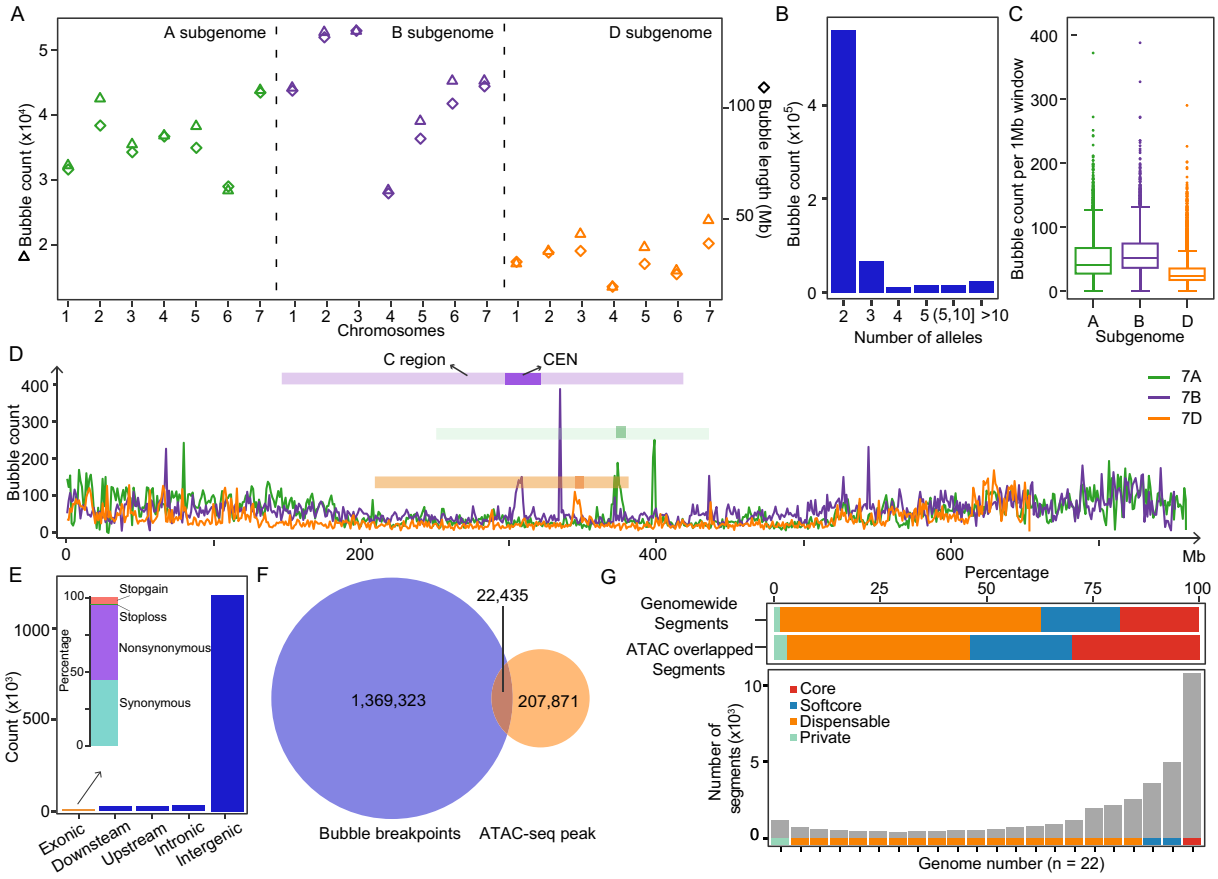

Fig. 3. Pangenome graph metrics. **A**: Total count and length of bubbles across 21 chromosomes. **B**: Allelic diversity of bubbles. The horizontal axis indicates the number of paths for each bubble. **C**: Bubble count per 1 Mb windows through the whole genome. Subgenomes are indicated by different colors. **D**: Distribution of bubbles for chromosomes 7A, 7B, and 7D. Dark- and light-colored bars above represent chromosomal segment C and the precise location of centromere (CEN), respectively. **E**: The number of bubble breakpoints overlapping with different genomic features. **F**: Venn diagram showing the extent of overlapping between bubble breakpoints and detected ATAC-seq peaks. **G**: Comparison of segments component for the whole genome and that overlapped with ATAC-seq peaks. Segments are divided into core, softcore, dispensable, and private according to frequency, the same as the criteria in gene-based pangenome.

different among SV types (Fig. 4B). The proportion of 100 bp–500 bp range was the highest for insertion (INS) (56.7%), deletion (DEL) (40.3%), and tandem repeat (TDM) (32.7%). 1 kb–10 kb range was dominant for duplicated region (DUP), inverted duplicated region (INVDP), inverted translocated region (INVTR), and translocated region (TRANS), ranging from 68.9% to 76.6%. The percentages of SVs over 50 kb were much higher for inversion (INV) and NOTAL compared to other SV types. Without regard to SV type, 87.2% of SVs were smaller than 10 kb in total (Fig. S19).

Similarly, patterns of length distribution were not exactly the same for different SV types. Copy gain (CPG) was primarily concentrated in ~200 bp, 1 kb, 1.8 kb and 2.6 kb, while copy loss (CPL) exhibit peaks at ~8.8 kb and 12.4 kb (Fig. 4C). Interestingly, the length distributions of INS and DEL were similar (Fig. 4C), characterized by four main peaks at about 100 bp, 4.2 kb, 8.8 kb and 14.2 kb, one of which consistent with CPL peak (8.8 kb) (Fig. 4C).

When compared with the graph bubbles, incorporating multiple SVs missing from the linear reference genome, we found 606,387 (30.65%) of the SVs overlapped with a graph bubble. 390,621 (56.13%) of bubbles contained at least one SV reciprocally (Fig. 4D). As for different SV types, the percentages varied greatly from 20.6% for INVDP to 84.1% for DEL, some of which were even lower than the total SV set (Fig. 4E). For real insight into influencing factors, we checked every step carefully. Minigraph only considers SVs of 100 bp–100 kb in length and ignores SVs in alignments shorter than 100 kb in the procedure of graph construction (Li et al., 2020). Nevertheless, there were 331,247 (16.7%) SVs with a length shorter than 100 bp and 39,388 (2.0%) SVs exceeding 100 kb in length. We speculated that length restrictions considerably resulted in the ineffective overlapping between SVs and bubbles.

We next turned our attention to SV distribution across the whole genome. Chromosomal zones exhibited a considerable difference in SV density, R1/R3 > R2a/R2b > C (Fig. 5A). This pattern was consistent with the distribution of other variations in wheat (IWGSC, 2018). However, SV density for windows in CENs was not low. Digging further, we found striking SV peaks around CENs (Fig. S20; Table S7), consistent with a recent study on wheat pangenome (Jiao et al., 2024). The SVs within CENs were mainly HDR, NOTAL, and CPG, accounting for 64.0%. The proportion of HDR was reduced from 48.4% (genome-wide) to 21.3% (CEN), while it increased from 9.3 % to 21.1% for CPG (Fig. 5B). Besides, the length distribution of CEN SVs was also different from that of the whole genome (Fig. 5C). We identified 497 SV hotspots stretching across ~948 Mb of the graph (Fig. 5D; Table S8). Of these hotspots, 43 were in C region, 122 in R2a/R2b region, and 260 in R1/R3 region.

Regional graph of wheat functional genes

To detect the trait-related SVs with potential phenotypic impact without accessible measured phenotypes of agronomic traits, which

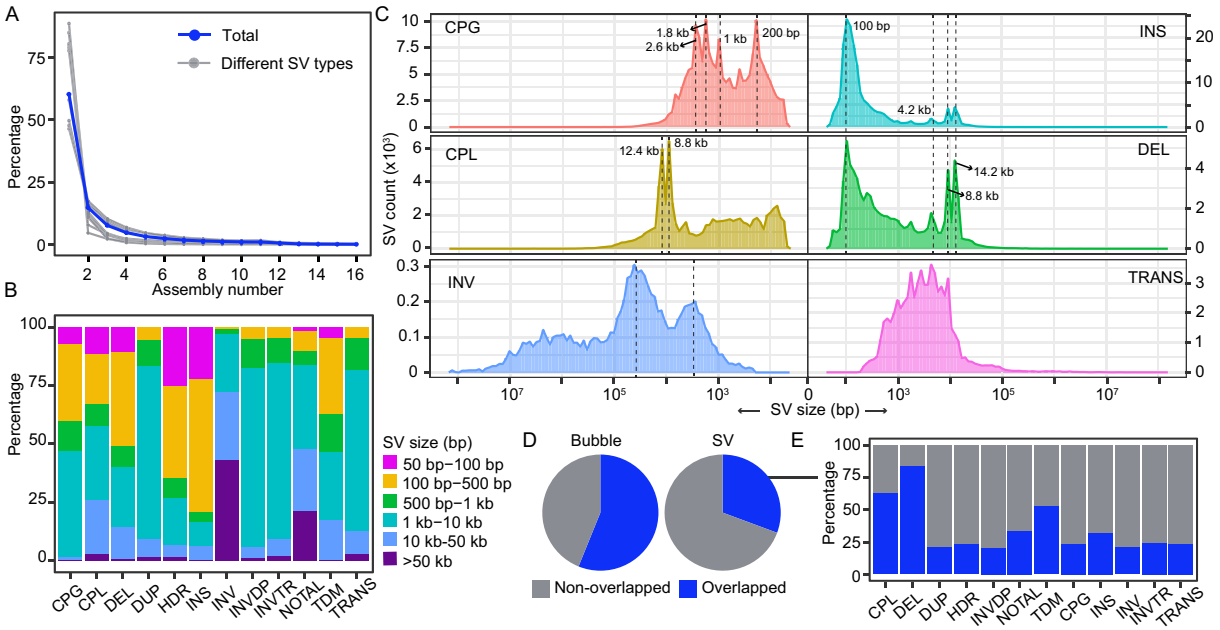

Fig. 4. Characterization of SVs. **A**: Frequency distribution of SVs across 16 chromosome-level genomes. **B**: The percentage of different length ranges for different SV types. **C**: Length distribution of different SVs. Six types of SVs were displayed: CPG for copy gain, CPL for copy loss, INS for insertion, DEL for deletion, INV for inversion, and TRANS for translocations. The vertical dashed lines indicate peaks, and their approximated locations are marked alongside. **D**: Validation of SVs using graph bubbles. Grey and navy colorsrepresent private and overlapped proportions, respectively. **E**: Percentages of SV overlapped with bubble for different SV types. For each type, the percentage was calculated as follows: overlapped count/ total count. SVs, structural variations.

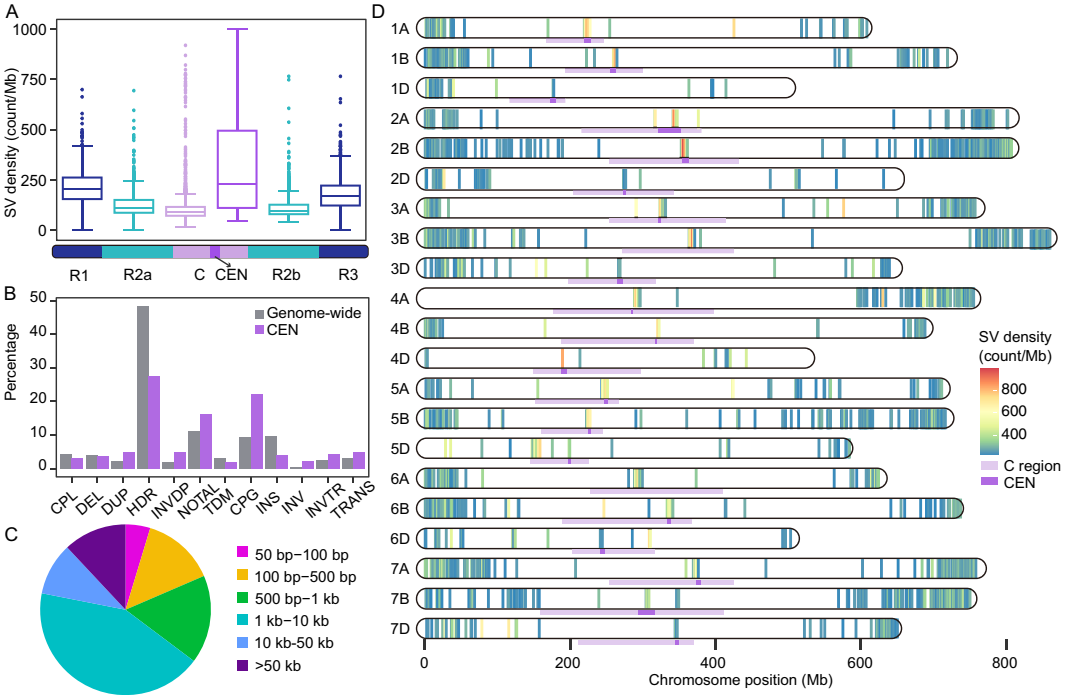

Fig. 5. Distribution of SVs. **A**: SV density of different chromosomal segments. Chromosomes were divided into five zones, R1, R2a, C, R2b, and R3, as inferred in the previous study (IWGSC, 2018). CEN represents centromere. Dimensions of chromosomal segments are displayed below the boxplot. Windows in CEN were excluded from the chromosomal "C" segment in the analysis. **B**: Comparison of SV composition on genome-wide and CEN. **C**: Percentage of SVs in different length ranges within CENs. **D**: Genome-wide distribution of identified SV hotspots. Coloring indicates the count of SVs in each non-overlapping 1 Mb window. Chromosomal C region and the precise position of CEN are highlighted by light and dark purple bars, respectively. SVs, structural variations; CEN, centromere.

were essential for GWAS and QTL mapping, we took advantage of the known growth habit information (Fig. 6A). We divided 22 accessions into two groups, winter (n = 10, 9 winter and 1 semi-winter accessions) and spring (n = 12, 11 spring and 1 facultative spring accessions), and then calculated the relative frequency differences (RFD) of individual segment in our graph between two growth habit

groups to investigate genomic regions and SVs controlling wheat growth habit (Fig. 6B). We used the $P < 0.0001$ (Z test) absolute RFD as the threshold for identifying "Spring-Winter" highly differentiated SVs. A total of 13,805 candidate segments (absolute RFD $\geq$0.57) were identified. We compared these outliers with previously mapped photoperiod and vernalization genes genetically controlling growth habits in wheat, including *Vrn* and *Ppd* series genes (Kamran et al., 2014; Hyles et al., 2020). However, the reported CNVs of *VRN-A1* (Jiao et al., 2024) among different ecotypes were not included. We further examined the copy numbers of *VRN-A1* in assemblies used in this study through pairwise alignment. A ~494 kb segment covering two copies of *VRN-A1* presented as PAVs, showing an inapparent difference in frequency between spring and winter groups. We speculated that the quite different set of samples and assembly quality probably resulted in the inconsistent findings.

Then, we focused on the segments showing much higher absolute RFD. We identified 2769 segments with absolute RFD exceeding 0.7 and 78 segments exceeding 0.9 (Fig. 6B and 6C), of which 11 (11/78) located in ATAC peaks. Among them, a 218-bp segment on chromosome 5B (596,657,558–596,657,775 bp) were present in all accessions within the spring group, while absent in 90% (9/10) accessions within the winter group, even in the accessions with 5B/7B translocation (Figs. 6C, 6D, S21, S22). The only one exception was a semi-winter wheat (Aikang58) (Fig. 6C). This variation was also verified in SV dataset. There exists a malate dehydrogenase (*MDH*) gene (chr5B:596,661,367–596,665,015, TraesFLD5B01G448200) downstream this variation according to the genome annotation of Fielder (Fig. 6D). *MDHs* play crucial roles in various cellular processes, such as carbon fixation, nitrogen metabolism, plant growth, and development (Baird et al., 2024). Two copies of *MDH* gene were localized on chromosome 1 and 5 across the three subgenomes of hexaploid wheat, *MDH1* and *MDH2*, which are cytoplasm targeted and mitochondrial targeted, respectively (Rangan et al., 2016). Transcriptomic and molecular evidence suggested that *MDH2* is probably involved in C4 photosynthesis in wheat grain development (Monson and Sage, 1999; Rangan et al., 2016).

To further evaluate the power of our graph-based pangenome in capturing genetic variations, we extracted and investigated the subgraph structures of 335 genes involved in more than 30 traits of wheat (Table S9). This systematical summarize of wheat functional genes was mainly supported by a previous review (Rasheed et al., 2024) and our long-term accumulation of wheat genomic research. We found that 258 of 319 genes (80.9%) being successfully positioned in our graph were highly conserved, showing a monoallelic state. Reduced plant height (*Rht*) genes, including *Rht1*, *Rht8*, and *Rht24*, were typical examples (Figs. S23–S25). Only one copy of *Rht8* contained SVs (Fig. S24).

Despite the high proportion of genes with plain graph, there were still numerous genes located inside of the complex and multiallelic regions of the graph (Figs. 6E, 6F, and S25). For instance, *PPD-D1* (chr2D:38,740,766–38,744,404), one of the primary photoperiod response genes for wheat (Beales et al., 2007), overlapped with a complex bubble made up of five segments, comprising three alternative paths (Fig. 6E). The 2089-bp segment (s7) upstream of *PPD-*

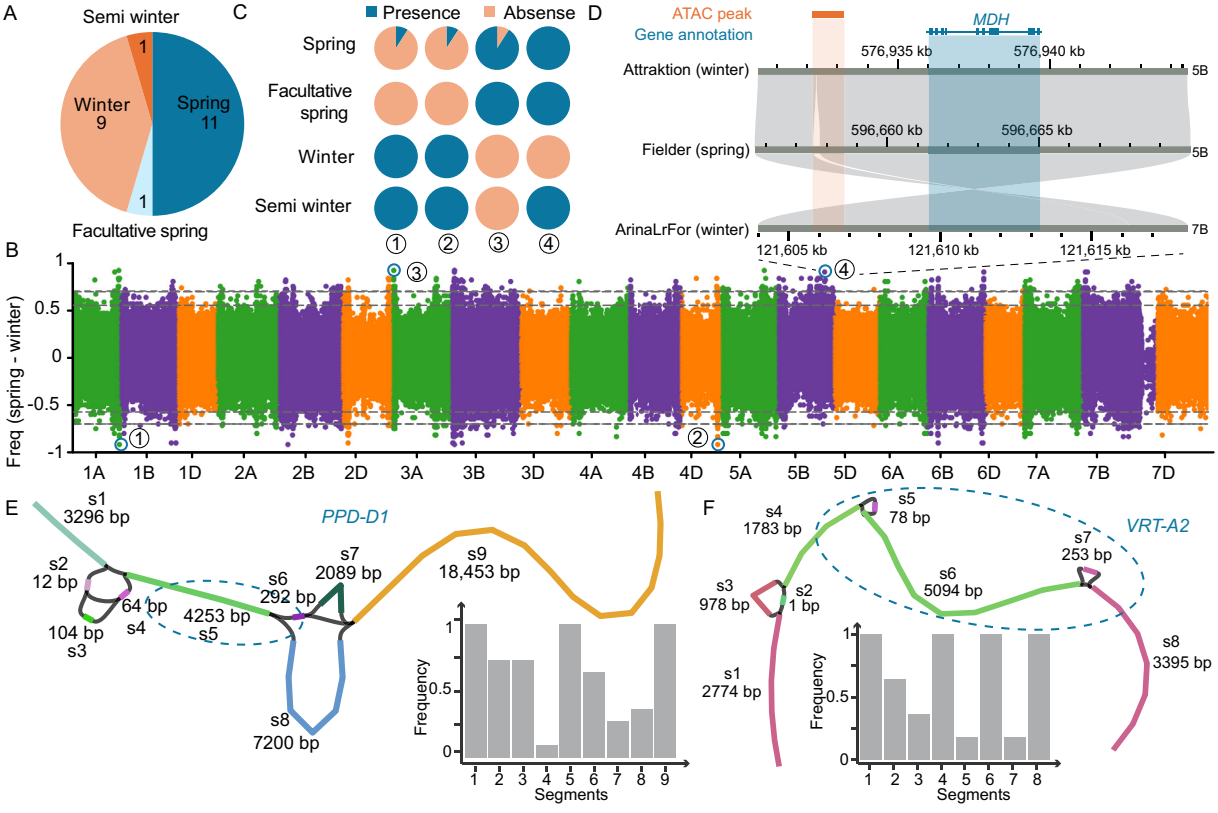

Fig. 6. Genome-wide screening and annotation of functional structural variations. **A**: Growth habit of 22 wheat assemblies. **B**: Genome-wide screening of segments showing significant frequency differences between spring and winter groups. The horizontal dashed grey lines indicate the threshold ($P < 0.0001$, Z test) with absolute RFD $\geq$ 0.567 and RFD $\geq$ 0.7. **C**: Frequency spectrum at four focused segments labeled in **B**. **D**: A 218-bp regulatory deletion upstream of *MDH* gene on chromosome 5B. The light orange and blue rectangle indicate the location of the ATAC-seq peak and *MDH* gene, respectively. The gene structure of *MDH* is displayed on the top. **E** and **F**: Visualization of structures surrounding functional genes with complex local graphs, **E** for *PPD-D1* and **F** for *VRT-A2*. The colored lines showed the paths of each segment. The blue dashed circles indicated the approximate position of gene bodies. The length and frequency of each segment were provided alongside the graphs.

D1 was identified as a ~2-kb deletion (varieties vs. CS) previously (Beales et al., 2007; Guo et al., 2010). Moreover, we detected a novel path covering a ~7-kb substitute allele (s8) (Fig. 6E). Next, we traced this segment to investigate the subgraph structure in each assembly. s8 was present in 8 of 22 assemblies. *VRT-A2*, driving glumes and grain elongation in wheat (Liu et al., 2021), included two biallelic bubbles, which were relatively low-complexity region (Fig. 6F). These findings indicated the potential applications of our graph in identification of more novel loci associated with important traits, especially for complex SVs that are absent in the linear reference genome.

Discussion

In this study, we constructed a gene-based and SV-based pangenome by integrating 22 assemblies of different hexaploid wheat lines. Our graph length is 1.14 times greater than the CS2.1 reference (Zhu et al., 2021). Moreover, we provided a comprehensive SV dataset and further identified a series of SVs with potential in phenotypic variations.

Wheat pangenome still needs to be improved

Although we have tried to integrate all the wheat assemblies available at present, we still need to admit that the graph constructed in this study was not the best. On the one hand, there was a large scale of sequences in the backbone genome and other assemblies that not aligned to each other (Fig. S16), resulting in a lack of potential location in graph. Most of these NOTAL SVs have not been integrated in our graph (Fig. 4E). Flanking sequences and Hi-C data can be applied to determine the physical coordinates of these NOTAL sequences (Tian et al., 2020), enable the integration of these SVs to the present graph. In addition, the assemblies using short-reads probably affected SV detection, potentially leading to false-positive identifications.

On the other hand, introducing more assemblies and long-read sequencing of diverse wheat landraces and their wild relatives will capture more comprehensive genetic variations and further improve the graph representativeness of this important crop species. Nevertheless, the number of inclusive genomes cannot be increased without restriction, since the paths increased exponentially with the increase of included variations, which would bring great challenge for computing resources and mapping accuracy while performing population-scale map-to-pan (Consortium, 2016; Outten and Warren, 2021). Objectively speaking, we have accomplished an important step forward in the wheat pan-genomic research, although it still needs improvement in graph accuracy and representativeness.

Comparison of gene families and graph segments component to existing wheat pangenome

As is described above, the frequency distribution of gene families in our gene-based graph slightly deviated from U-shape pattern, which usually is pictured in pangenome analyses for other species (Collins and Higgs, 2012; Liu et al., 2020; Chen et al., 2023; Shi et al., 2024; Zhang et al., 2024). We supposed such a phenomenon was associated with gene projection strategies. The gene projection of most assemblies (16/22), including 10+ Wheat Genomes ($n = 15$) and Fielder, were based on high-confidence gene models in Chinese Spring using alignment (IWGSC, 2018; Walkowiak et al., 2020; Sato et al., 2021), limiting the investigation of private genes in each genome. The gene projections of Kariega (Athiyannan et al., 2022) and Aikang58 (Jia et al., 2023) were only supported by RNA-seq data. There were only two accessions, Zang1817 (Guo et al., 2020) and Renan (Aury et al., 2022), which used a comprehensive strategy by combining *Ab-initio* prediction, homology-based prediction, and RNA-seq-based prediction. Such non-uniform annotation methods inevitably influenced the number and quality of gene models.

Furthermore, the percentage of core genes (23.19%) was quite low as compared with that reported in previous studies: 64.3% for 18 cultivars mainly from Australian (16/18) (Montenegro et al., 2017), 65.66% for 17 cultivars mainly from China (16/17) (Jiao et al., 2024). Similarly, the percentage of core segments was lower (9.14%) than that estimated in a wheat pangenome (19%) (Bayer et al., 2022), which included 15 high-quality bread wheat genome assemblies fully covered in our study. This percentage showed a wide range among different species, 40.2% in *Arabidopsis thaliana* (Lian et al., 2024), 37.9%−46.5% in rice (Wang et al., 2018), and 64.3% in soybean (Liu et al., 2020). For one certain species, the percentage also differed when using different individuals. We speculated that these variations likely stem from the genetic diversity of the selected wheat assemblies. In summary, it was difficult to determine the real landscape of gene PAVs in wheat using available data at present. Improving the genome annotation quality by taking advantage of federated gene prediction strategy would be beneficial to decipher it.

Enrichment of SVs in centromere

CEN, a constitutive structure of the chromosomes, plays an important role in chromosomal pairing during cell division (Kursel and Malik, 2016). For hexaploidy wheat, this paring process was particularly challenging. A series of explorations on wheat CEN positioning, dynamics, and evolution had been performed (Su et al., 2019; Wang et al., 2022; Ma et al., 2023; Zhao et al., 2023a). To some extent, CENs were stable in wheat; large centromeric ancestral haplotypes group (centAHG) spanning from 80 to 250 Mb acted as backbone of wheat chromosomes (Wang et al., 2022). Whereas multilevel evidence also supported great dynamics of wheat CENs, the genetic diversity of inter-centAHG highly increased than intra-centAHG (Wang et al., 2022). Moreover, obvious physical shifts of CENs were detected in Aikang58 compared with CS (Zhao et al., 2023a). Another study using wheat 10+ lines further captured a high degree of sequences, positioning, and epigenetic variations in CEN (Ma et al., 2023). But there are few studies of SV landscape in wheat CEN and their relationship with CEN evolution at the population level in wheat.

In this study, we systematically identified SVs across the whole genome using 22 wheat lines and observed remarkable increasement of bubble and SV density in almost all of wheat CENs (Figs. 3D, 5D, S11, S20). Interestingly, such SV enrichment also remained on chromosomes involving large-scale pericentric inversions or translocations, suggesting the persistence of these SVs prior to the variety development. The highly structural dynamics of CEN has been increasingly reported in other species, such as *Arabidopsis thaliana* (Lian et al., 2024) and *Brachypodium* (Chen et al., 2024). For *Brachypodium* genus, high degrees of sequence and epigenetic variations of CEN between different lines were also revealed (Chen et al., 2024). In human, about a 4.1-fold increase in SNPs appeared when compared two sets of human centromeres (Logsdon et al., 2024). To enable the comparison between SVs and SNPs/Indels in this study, we revisited our previous analyses (Huang et al., 2023), but there was indeed no remarkable increase of SNPs/Indels in CENs. Based on these findings, we speculated those larger SVs, rather than small SNPs/Indels, probably play a vital role in the high robustness for CEN function maintenance of wheat and ensuring the stability of karyotypes. Our findings promoted the research on CEN plasticity and stability of wheat. But it was unclear whether these SVs in CENs were associated with phenotypic variations of wheat different lines. Their potential impact on phenotype and breeding process needs further study.

Materials and methods

Data sources

All the genome sequences and ATAC-seq data in this study were obtained from public deposition under accessions (Table S1). The completeness of whole genome sequence and protein sequence was assessed by the BUSCO (v.4.0) (Simão et al., 2015), respectively.

Repeat annotation

We used EDTA (Extensive *de-novo* TE Annotator) (Ou et al., 2019), a single-package approach to annotate different types of TEs in all 22 assemblies. TEs were mainly classified into LRT (Copia, Gypsy, and TRIM), TIR (CACTA, hAT, Mutator, PIF Harbinger, and Tc1 Mariner), nonLTR (LINE element), nonTIR (LINE element), and simple repeat region.

Genome alignment

To achieve the accurate alignment among such complex and huge assemblies, we used complementary approaches that can enable us to explore SVs both between chromosomes (interchromosomal) and within chromosomes (intrachromosomal). Using the Fielder genome as the reference, the other 16 chromosome-level assemblies were aligned to it genome by genome using AnchorWave (Song et al., 2022) and chromosome by chromosome using MUMmer (Marçais et al., 2018). AnchorwWave has a great advantage in running time and memory, but MAF (Multiple Alignment Format) output only recorded the alignment of anchors, which were gene sequences in this study. Therefore, the results of AnchorWave were mainly used to identify the large-scale interchromosomal SVs. The result of MUMmer showing precise alignment positions was used to detect smaller intrachromosomal SVs.

Construction of the protein-coding gene-based pangenome

First, we extracted protein sequences for projected genes. For genes with multiple transcripts, only the longest transcript was retained. ArinaLrFor, lacking genome annotation, and Renan and Kariega, showing low BUSCO, were excluded from 22 assemblies in this section. Then, we used OrthoFinder (Emms and Kelly, 2019) to identify gene families. The simulation of pan- and core-genome size (Fig. 2A), calculation of core, softcore, dispensable, and private gene families, and their distribution in each accession (Fig. 2B–2D) were performed using custom scripts.

Construction of pseudochromosomes and graph-based pangenome

Five of the genomes in this study were assembled to the scaffold level (Cadenza, Paragon, Robigus, Claire, and Weebill), which were incompatible with minigraph (Li et al., 2020). We used RagTag (Alonge et al., 2022), a homology-based toolset for assembly scaffolding and patching, to establish chromosome-scale genomes of these five scaffold assemblies. Then, we used minigraph to construct wheat graph pangenome with Fielder as reference genome. The output graph was in Graphical Fragment Assembly (GFA) format, which describes constituent segments (lines starting with "S") and links between them (lines starting with "L") (Li et al., 2020). The raw graph contained 2,618,041 segments.

To improve the quality of our pangenome, we realigned individual genomes to the graph using minigraph (Li et al., 2020), respectively. According to the reported depth of coverage in output Graphical mApping Format (GAF) file, only the segments supported by at least one assembly (coverage > 0.9) were retained. After being filtered, there were 2,580,200 segments retained in the final graph pangenome (Fig. S10). Then, we used "gfatools bubble" to detect the SV sites of graph. Gfatools is available at https://github.com/lh3/gfatools. The minigraph SVs output is a BED-like file, giving the position, source, list of the segments, and all possible paths of a bubble variation. We counted the bubbles with different numbers of paths/alleles to estimate wheat genetic diversity (Fig. 3B). To explore the bubble distribution across the whole genome, we calculated the number of bubbles in non-overlapping 1 Mb windows (Fig. S11).

Positioning of centromeres in Fielder genome

The precise positions for centromeres and chromosomal structural partitioning (R1, R2a, C, R2b, R3) have been defined in IWGSC RefSeq v1.0 (IWGSC, 2018). To determine the boundary locations of these segments in our graph, we truncated sequences surrounding the boundaries from IWGSC RefSeq v1.0 and performed sequence alignment between IWGSC RefSeq v1.0 and Fielder genome (Sato et al., 2021). This approach was reasonable since core centromere repeats are homogeneous and dominant in different wheat lines (Ma et al., 2023). Except for the locations of R2b/R3 boundary on 5B and C/R2b boundary on 6B, all the boundary locations and centromere positions were successfully resolved (Table S7).

Identification of ATAC peak

Raw sequencing data of ATAC-seq were downloaded from NCBI Sequence Read Archive under accession number PRJNA779945 (Pei et al., 2023b) and PRJNA886977 (Pei et al., 2023a), from European Nucleotide Archive under accession number PRJEB29868(Lu et al., 2020), and from Genome Sequence Archive (https://bigd.big.ac.cn/gsa) under accession number PRJCA008382 (Zhao et al., 2023b). SRA files were converted to FASTQ format using "fastq-dump" function in SRA Toolkit. Adapter trimming and quality filtering of raw reads were performed using fastp (Chen et al., 2018). The processed reads were then mapped to the IWGSC RefSeq v1.0 using Bowtie2 (Langmead and Salzberg, 2012). Mapped reads were sorted using samtools (Danecek et al., 2021), and redundant reads and PCR duplicates were marked using Picard (Wysokar et al., 2014).

Peak calling was performed using MACS2 with the following parameters "-g 14e9 –nomodel -q 0.05" (Feng et al., 2012). To improve the accuracy, duplicated reads were not considered ("–keep-dup = 1"). To mitigate experimental noise from different projects, only peaks present in at least 20% of the samples were retained. Finally, we detected a total of 230,307 ATAC peaks, covering 185,338,412 bp (1.30%) sequence of the genome. The average length of peak was ~804 bp.

Identification of SVs

Based on the results of MUMmer above, we used identified SyRI (Goel et al., 2019) to identify SVs and used plotsr (Goel and Schneeberger, 2022) to visualize the synteny and SVs between genomes (Figs. S13–S15). The SVs were classified into 12 types in SyRI output, including CPL, DEL, DUP, HDR, INVDP, NOTAL, TDM, GPG, INS, INV, INVTR, and TRANS (Fig. S16; Table S6). In consideration of the effect of scaffold-level assemblies with low Contig N50 values on SV detection, only the chromosome-level assemblies were used in SVs identification. We classified these variations into three main types according to previous relevant studies (Qin et al., 2021): PAVs (Presence/Absence Variations), inversions, and translocations. CPL, DEL, DUP (copyloss), INVDP (copyloss), and the Fielder sequence in HDR, NOTAL, and TDM were converted as Absence SVs in a given

accession. CPG, INS, DUP (copygain), INVDP (copygain), and the query sequences in HDR, NOTAL, and TDM were converted as Presence SVs. INVs were regarded as inversion SVs. TRANS and INVTR were both regarded as translocation SVs. After filtering out all SVs less than 50 bp in length, a total of 1,1978,221 SVs were detected in at least one genome as compared with Fielder genome.

SVs merge and hotspot identification

The SVs identified by SyRI above were merged into a non-redundant call set using SURVIVOR (Jeffares et al., 2017). The command is as follows: SURVIVOR merge input_filelist 1000 1 0 0 0 50, in which "1000" means SVs within 1 kb between two assemblies were merged, and "50" means only SVs with lengths larger than 50 bp were considered. To improve the accuracy, different types of SVs were merged separately and then integrated into a final set. INS variations that cannot be merged by SURVIVOR were merged using the merge tool of BEDTools (Quinlan and Hall, 2010). It should be specially explained that NOTAL variations in query genomes, which were absent in Fielder, do not have corresponding coordinates in Fielder genome. Therefore, these NOTAL variations cannot be merged to a common coordinate to enable the exploration on population scale. In total, there were 334,035 such NOTAL SVs.

Using the non-redundant call set above, we investigated SVs distribution in non-overlapping 1 Mb sliding windows as the calculation for bubbles. All windows were ranked in descending order according to the number of SVs within the window. We defined the top 1 % of all windows with the highest density of SVs as SV hotspots, then merged hotspot windows separated by a distance of $\leq$1 Mb as "hotspot regions". We obtained 1440 hotspot windows and 497 hotspot regions (Fig. 5D).

Differentiated SVs between spring and winter wheat groups

We divided 22 accessions in this study into two groups according to growing habits. "Spring" group ($n = 12$) includes 11 spring and 1 facultative spring accessions. "Winter" group ($n = 10$) includes 9 winter and 1 semi-winter accessions (Table S1). To identify the differentiated SVs between spring and winter groups, we calculated the frequency of each segment in our graph pangenome for the two groups, respectively. Next, we used the relative frequency difference to measure the differentiation of these segments. The formula was as follows:

$RFD = F_{spring} - F_{winter}$where F_{spring} and F_{winter} represent the frequency of individual segments in spring and winter groups, respectively. We calculated the P values according to Z-transformed RFD, and the segments with $P < 0.001$ were defined as potential regions associated with growth habit regulation. This threshold is much higher than the highest 1 % ($\geq$0.50) and genome-wide mean absolute RFD ($\geq$0.12).

Local graph investigation of functional genes

Currently discovered genes associated with various traits, including yield, disease resistance, and adaptability, were used to locate the corresponding copies in our graph. Since most of these genes were reported in IWGSC RefSeq v1.0 coordinate, we truncated genic sequences from IWGSC RefSeq v1.0 and then aligned them to Fielder (the backbone of graph pangenome) using BLAST to determine their position in our graph. Three hundred and nineteen of the 335 functional genes found appropriate hits showing high sequence identity and complete coverage. Then, we used "Gfatools view" to extract subgraphs overlapping functional genes as well as other candidate regions and used Bandage to visualize the bubbles/nodes of subgraphs (Wick et al., 2015).

CRediT authorship contribution statement

Hong Cheng: Investigation, Bioinformatics analyses, Data curation, Conceptualization, Visualization, Writing - Original draft. **Lingpeng Kong**: Methodology, Investigation, Data curation. **Kun Zhu**: Methodology, Visualization. **Xiuli Li**: Visualization. **Hang Zhao**, **Yanwen Zhang**: Visualization. **Weidong Ning** and **Mei Jiang**: Data curation. **Bo Song**: Writing - Review & Editing. **Shifeng Cheng**: Supervision, Project administration, Funding acquisition.

Data availability

The sequence and annotation of published genome used in this study are all downloaded from public database (Table S1). All data are freely available http://wheatpgdb.cn/.

Code availability

The source code and scripts used in this study have been deposited in GitHub (https://github.com/Chenghong412/wheat_pangenome).

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2023YFF1000100 and 2023YFA0914601) and the Special Funds for Science Technology Innovation and Industrial Development of Shenzhen Dapeng New District (PT202101-01).

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jgg.2025.03.015.

References

Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., Wang, X., Lippman, Z.B., Schatz, M.C., Soyk, S., 2022. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol. 23, 258.

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D., et al., 2020. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell 182, 145–161.

Athiyannan, N., Abrouk, M., Boshoff, W.H.P., Cauet, S., Rodde, N., Kudrna, D., Mohammed, N., Bettgenhaeuser, J., Botha, K.S., Derman, S.S., et al., 2022. Long-read genome sequencing of bread wheat facilitates disease resistance gene cloning. Nat. Genet. 54, 227–231.

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

Baird, L.M., Berndsen, C.E., Monroe, J.D., 2024. Malate dehydrogenase in plants: evolution, structure, and a myriad of functions. Essays Biochem. 68, 221–233.

Bayer, P.E., Golicz, A.A., Scheben, A., Batley, J., Edwards, D., 2020. Plant pangenomes are the new reference. Nat. Plants 6, 914–920.

Bayer, P.E., Petereit, J., Durant, E., Monat, C., Rouard, M., Hu, H., Chapman, B., Li, C., Cheng, S., Batley, J., et al., 2022. Wheat Panache: a pangenome graph database representing presence-absence variation across sixteen bread wheat genomes. Plant Genome 15, e20221.

Beales, J., Turner, A., Griffiths, S., Snape, J.W., Laurie, D.A., 2007. A *Pseudo-response regulator* is misexpressed in the photoperiod insensitive *Ppd-D1a* mutant of wheat (*Triticum aestivum* L.). Theor. Appl. Genet. 115, 721–733.

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G.L., D'Amore, R., Allen, A.M., McKenzie, N., Kramer, M., Kerhornou, A., Bolser, D., et al., 2012. Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 491, 705–710.

Cavanagh, C.R., Chao, S., Wang, S., Huang, B.E., Stephen, S., Kiani, S., Forrest, K., Saintenac, C., Brown-Guedira, G.L., Akhunova, A., 2013. Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proc. Natl. Acad. Sci. U. S. A. 110, 8057–8062.

Chen, C., Wu, S., Sun, Y., Zhou, J., Chen, Y., Zhang, J., Birchler, J.A., Han, F., Yang, N., Su, H., 2024. Three near-complete genome assemblies reveal substantial centromere dynamics from diploid to tetraploid in *Brachypodium* genus. Genome Biol. 25, 63.

Chen, J., Liu, Y., Liu, M., Guo, W., Wang, Y., He, Q., Chen, W., Liao, Y., Zhang, W., Gao, Y., et al., 2023. Pangenome analysis reveals genomic variations associated with domestication traits in broomcorn millet. Nat. Genet. 55, 2243–2254.

Chen, S., Zhou, Y., Chen, Y., Gu, J., 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884–i890.

Cheng, H., Liu, J., Wen, J., Nie, X., Xu, L., Chen, N., Li, Z., Wang, Q., Zheng, Z., Li, M., et al., 2019. Frequent intra- and inter-species introgression shapes the landscape of genetic variation in bread wheat. Genome Biol. 20, 136.

Cheng, S., Feng, C., Wingen, L.U., Cheng, H., Riche, A.B., Jiang, M., Leverington-Waite, M., Huang, Z., Collier, S., Orford, S., et al., 2024. Harnessing landrace diversity empowers wheat breeding. Nature 632, 823–831.

Clavijo, B.J., Venturini, L., Schudoma, C., Accinelli, G.G., Kaithakottil, G., Wright, J., Borrill, P., Kettleborough, G., Heavens, D., Chapman, H., et al., 2017. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. Genome Res. 27, 885–896.

Collins, R.E., Higgs, P.G., 2012. Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. Mol. Biol. Evol. 29, 3413–3425.

Consortium, T.C.P.-G., 2016. Computational pan-genomics: status, promises and challenges. Brief. Bioinf. 19, 118–135.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al., 2021. Twelve years of samtools and bcftools. GigaScience 10, giab008.

Eizenga, J.M., Novak, A.M., Sibbesen, J.A., Heumos, S., Ghaffaari, A., Hickey, G., Chang, X., Seaman, J.D., Rounthwaite, R., Ebler, J., et al., 2020. Pangenome graphs. Annu. Rev. Genom. Hum. Genet. 21, 139–162.

Emms, D.M., Kelly, S., 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20, 238.

FAO, 2022. Agricultural Production Statistics 2000–2020, 41. FAOSTAT Analytical Brief Series, Rome.

Feng, J., Liu, T., Qin, B., Zhang, Y., Liu, X.S., 2012. Identifying CHIP-seq enrichment using MACS. Nat. Protoc. 7, 1728–1740.

Goel, M., Schneeberger, K., 2022. plotsr: visualizing structural similarities and rearrangements between multiple genomes. Bioinformatics 38, 2922–2926.

Goel, M., Sun, H., Jiao, W.B., Schneeberger, K., 2019. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol. 20, 277.

Golicz, A.A., Bayer, P.E., Barker, G.C., Edger, P.P., Kim, H.R., Martinez, P.A., Chan, C.K.K., Severnellis, A., Mccombie, W.R., Parkin, I.A.P., 2016. The pangenome of an agronomically important crop plant *Brassica oleracea*. Nat. Commun. 7, 13390.

Guo, W., Xin, M., Wang, Z., Yao, Y., Hu, Z., Song, W., Yu, K., Chen, Y., Wang, X., Guan, P., et al., 2020. Origin and adaptation to high altitude of Tibetan semi-wild wheat. Nat. Commun. 11, 5085.

Guo, Z., Song, Y., Zhou, R., Ren, Z., Jia, J., 2010. Discovery, evaluation and distribution of haplotypes of the wheat *Ppd-D1* gene. New Phytol. 185, 841–851.

Huang, X., Rymbekova, A., Dolgova, O., Lao, O., Kuhlwilm, M., 2023. Harnessing deep learning for population genetic inference. Nat. Rev. Genet. 25, 61–78.

Hubner, S., Bercovich, N., Todesco, M., Mandel, J.R., Odenheimer, J., Ziegler, E., Lee, J.S., Baute, G.J., Owens, G.L., Grassa, C.J., et al., 2019. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. Nat. Plants 5, 54–62.

Hyles, J., Bloomfield, M.T., Hunt, J.R., Trethowan, R.M., Trevaskis, B., 2020. Phenology and related traits for wheat adaptation. Heredity 125, 417–430.

IWGSC, 2014. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science 345, 1251788.

IWGSC, 2018. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361, eaar7191.

Jeffares, D.C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bahler, J., Sedlazeck, F.J., 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat. Commun. 8, 14061.

Jia, J., Zhao, G., Li, D., Wang, K., Kong, C., Deng, P., Yan, X., Zhang, X., Lu, Z., Xu, S., et al., 2023. Genome resources for the elite bread wheat cultivar Aikang 58 and mining of elite homeologous haplotypes for accelerating wheat improvement. Mol. Plant, 16, 1893–1910.

Jiao, C., Xie, X., Hao, C., Chen, L., Xie, Y., Garg, V., Zhao, L., Wang, Z., Zhang, Y., Li, T., et al., 2024. Pan-genome bridges wheat structural variations with habitat and breeding. Nature 637, 384–393.

Kale, S.M., Schulthess, A.W., Padmarasu, S., Boeven, P.H.G., Schacht, J., Himmelbach, A., Steuernagel, B., Wulff, B.B.H., Reif, J.C., Stein, N., et al., 2022. 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.

Kamran, A., Iqbal, M., Spaner, D., 2014. Flowering time in wheat (*Triticum aestivum* L.): a key factor for global adaptability. Euphytica 197, 1–26.

Kang, M., Wu, H., Liu, H., Liu, W., Zhu, M., Han, Y., Liu, W., Chen, C., Song, Y., Tan, L., et al., 2023. The pan-genome and local adaptation of *Arabidopsis thaliana*. Nat. Commun. 14, 6259.

Kursel, L.E., Malik, H.S., 2016. Centromeres. Curr. Biol. 26, R487–R490.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.

Li, H., Feng, X., Chu, C., 2020. The design and construction of reference pangenome graphs with minigraph. Genome Biol. 21, 265.

Li, H., Wang, S., Chai, S., Yang, Z., Zhang, Q., Xin, H., Xu, Y., Lin, S., Chen, X., Yao, Z., et al., 2022. Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. Nat. Commun. 13, 682.

Li, Y.H., Zhou, G., Ma, J., Jiang, W., Jin, L.G., Zhang, Z., Guo, Y., Zhang, J., Sui, Y., Zheng, L., et al., 2014. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat. Biotechnol. 32, 1045–1052.

Lian, Q., Huettel, B., Walkemeier, B., Mayjonade, B., Lopez-Roques, C., Gil, L., Roux, F., Schneeberger, K., Mercier, R., 2024. A pan-genome of 69 *Arabidopsis thaliana* accessions reveals a conserved genome structure throughout the global species range. Nat. Genet. 56, 982–991.

Liu, J., Chen, Z., Wang, Z., Zhang, Z., Xie, X., Wang, Z., Chai, L., Song, L., Cheng, X., Feng, M., et al., 2021. Ectopic expression of *VRT-A2* underlies the origin of *Triticum polonicum* and *Triticum petropavlovskyi* with long outer glumes and grains. Mol. Plant 14, 1472–1488.

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.A., Zhang, H., Liu, Z., Shi, M., et al., 2020. Pan-genome of wild and cultivated soybeans. Cell 182, 162–176.e113.

Logsdon, G.A., Rozanski, A.N., Ryabov, F., Potapova, T., Shepelev, V.A., Catacchio, C.R., Porubsky, D., Mao, Y., Yoo, D., Rautiainen, M., et al., 2024. The variation and evolution of complete human centromeres. Nature 629, 136–145.

Lu, F.-H., McKenzie, N., Gardiner, L.-J., Luo, M.-C., Hall, A., Bevan, M.W., 2020. Reduced chromatin accessibility underlies gene expression differences in homologous chromosome arms of diploid *Aegilops tauschii* and hexaploid wheat. GigaScience 9, giaa070.

Lyu, X., Xia, Y., Wang, C., Zhang, K., Deng, G., Shen, Q., Gao, W., Zhang, M., Liao, N., Ling, J., et al., 2023. Pan-genome analysis sheds light on structural variation-based dissection of agronomic traits in melon crops. Plant Physiol. 193, 1330–1348.

Ma, H., Ding, W., Chen, Y., Zhou, J., Chen, W., Lan, C., Mao, H., Li, Q., Yan, W., Su, H., 2023. Centromere plasticity with evolutionary conservation and divergence uncovered by wheat 10+ genomes. Mol. Biol. Evol. 40, msad176.

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., Zimin, A., 2018. MUMmer4: a fast and versatile genome alignment system. PLoS Comput. Biol. 14, e1005944.

Monson, R.K., Sage, R.F., 1999. C4 Plant Biology.

Montenegro, J.D., Golicz, A.A., Bayer, P.E., Hurgobin, B., Lee, H., Chan, C.K., Visendi, P., Lai, K., Doležel, J., Batley, J., 2017. The pangenome of hexaploid bread wheat. Plant J. 90, 1007–1013.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B., Elliott, T.A., Ware, D., Peterson, T., et al., 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 20, 275.

Outten, J., Warren, A., 2021. Methods and developments in graphical pangenomics. J. Indian Inst. Sci. 101, 485–498.

Pei, H., Li, Y., Liu, Y., Liu, P., Zhang, J., Ren, X., Lu, Z., 2023a. Chromatin accessibility landscapes revealed the subgenome-divergent regulation networks during wheat grain development. aBIOTECH 4, 8–19.

Pei, H., Teng, W., Gao, L., Gao, H., Ren, X., Liu, Y., Jia, J., Tong, Y., Wang, Y., Lu, Z., 2023b. Low-affinity spl binding sites contribute to subgenome expression divergence in allohexaploid wheat. Sci. China Life Sci. 66, 819–834.

Pont, C., Leroy, T., Seidel, M., Tondelli, A., Duchemin, W., Armisen, D., Lang, D., Bustos-Korts, D., Goue, N., Balfourier, F., et al., 2019. Tracing the ancestry of modern bread wheats. Nat. Genet. 51, 905–911.

Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., He, Q., Ou, S., Zhang, H., Li, X., et al., 2021. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. Cell 184, 3542–3558.

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842.

Rangan, P., Furtado, A., Henry, R.J., 2016. New evidence for grain specific C4 photosynthesis in wheat. Sci. Rep. 6, 31721.

Rasheed, A., Qayyum, H., Appels, R., 2024. Genome-informed discovery of genes and framework of functional genes in wheat. In: Appels, R., Eversole, K., Feuillet, C., Gallagher, D. (Eds.), The Wheat Genome. Springer International Publishing, Cham, pp. 165–186.

Sato, K., Abe, F., Mascher, M., Haberer, G., Gundlach, H., Spannagl, M., Shirasawa, K., Isobe, S., 2021. Chromosome-scale genome assembly of the transformation-amenable common wheat cultivar 'Fielder'. DNA Res. 28, dsab008.

Schreiber, M., Jayakodi, M., Stein, N., Mascher, M., 2024. Plant pangenomes for crop improvement, biodiversity and evolution. Nat. Rev. Genet. 25, 563–577.

Scott, M.F., Fradgley, N., Bentley, A.R., Brabbs, T., Corke, F., Gardner, K.A., Horsnell, R., Howell, P., Ladejobi, O., Mackay, I.J., et al., 2021. Limited haplotype

diversity underlies polygenic trait architecture across 70 years of wheat breeding. Genome Biol. 22, 137.
Shi, J., Tian, Z., Lai, J., Huang, X., 2023. Plant pan-genomics and its applications. Mol. Plant 16, 168–186.
Shi, T., Zhang, X., Hou, Y., Jia, C., Dan, X., Zhang, Y., Jiang, Y., Lai, Q., Feng, J., Feng, J., et al., 2024. The super-pangenome of populus unveils genomic facets for its adaptation and diversification in widespread forest trees. Mol. Plant 17, 725–746.
Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212.
Song, B., Marco-Sola, S., Moreto, M., Johnson, L., Buckler, E.S., Stitzer, M.C., 2022. AnchorWave: sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication. Proc. Natl. Acad. Sci. U. S. A. 119, e2113075119.
Su, H., Liu, Y., Liu, C., Shi, Q., Huang, Y., Han, F., 2019. Centromere satellite repeats have undergone rapid changes in polyploid wheat subgenomes. Plant Cell 31, 2035–2051.
Tian, X., Li, R., Fu, W., Li, Y., Wang, X., Li, M., Du, D., Tang, Q., Cai, Y., Long, Y., et al., 2020. Building a sequence map of the pig pan-genome from multiple de novo assemblies and Hi-C data. Sci. China Life Sci. 63, 750–763.
Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M.T., Brinton, J., Ramirez-Gonzalez, R.H., Kolodziej, M.C., Delorean, E., Thambugala, D., et al., 2020. Multiple wheat genomes reveal global variation in modern breeding. Nature 588, 277–283.
Wang, M., Wang, Y., Li, X., Zhang, Y., Chen, X., Liu, J., Qiua, Y., Wang, A., 2024. Integration of metabolomics and transcriptomics reveals the regulation mechanism of the phenylpropanoid biosynthesis pathway in insect resistance traits in *Solanum habrochaites*. Hortic. Res. 11, uhad277.
Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M., Zheng, T., Fuentes, R.R., Zhang, F., et al., 2018. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557, 43–49.
Wang, Z., Wang, W., Xie, X., Wang, Y., Yang, Z., Peng, H., Xin, M., Yao, Y., Hu, Z., Liu, J., et al., 2022. Dispersed emergence and protracted domestication of polyploid wheat uncovered by mosaic ancestral haploblock inference. Nat. Commun. 13, 3891.
Wick, R.R., Schultz, M.B., Zobel, J., Holt, K.E., 2015. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics 31, 3350–3352.
Wysokar, A., Tibbetts, K., McCown, M., Homer, N., Fennell, T., 2014. Picard: a set of tools for working with next generation sequencing data in BAM format. Retrieved Aug 2014 from. http://picardsourceforgenet.
Yan, H., Sun, M., Zhang, Z., Jin, Y., Zhang, A., Lin, C., Wu, B., He, M., Xu, B., Wang, J., et al., 2023. Pangenomic analysis identifies structural variation associated with heat tolerance in pearl millet. Nat. Genet. 55, 507–518.
Zhang, Z., Zhang, J., Kang, L., Qiu, X., Xu, S., Xu, J., Guo, Y., Niu, Z., Niu, B., Bi, A., et al., 2024. Structural variation discovery in wheat using PacBio high-fidelity sequencing. Plant J. 120, 678–689.
Zhang, Y., Zhao, M., Tan, J., Huang, M., Chu, X., Li, Y., Han, X., Fang, T., Tian, Y., Jarret, R., et al., 2024. Telomere-to-telomere *Citrullus* super-pangenome provides direction for watermelon breeding. Nat. Genet. 56, 1750–1761.
Zhao, J., Xie, Y., Kong, C., Lu, Z., Jia, H., Ma, Z., Zhang, Y., Cui, D., Ru, Z., Wang, Y., et al., 2023a. Centromere repositioning and shifts in wheat evolution. Plant Commun. 4, 100556.
Zhao, L., Yang, Y., Chen, J., Lin, X., Zhang, H., Wang, H., Wang, H., Bie, X., Jiang, J., Feng, X., et al., 2023b. Dynamic chromatin regulatory programs during embryogenesis of hexaploid wheat. Genome Biol. 24, 7.
Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., Wu, Y., Cheng, L., Fang, Y., Wu, K., et al., 2022. Graph pangenome captures missing heritability and empowers tomato breeding. Nature 606, 527–534.
Zhou, Y., Zhao, X., Li, Y., Xu, J., Bi, A., Kang, L., Xu, D., Chen, H., Wang, Y., Wang, Y.G., et al., 2020. *Triticum* population sequencing provides insights into wheat adaptation. Nat. Genet. 52, 1412–1422.
Zhu, T., Wang, L., Rimbert, H., Rodriguez, J.C., Deal, K.R., De Oliveira, R., Choulet, F., Keeble-Gagnere, G., Tibbits, J., Rogers, J., et al., 2021. Optical maps refine the bread wheat *Triticum aestivum* cv. Chinese Spring genome assembly. Plant J. 107, 303–314.
Zimin, A.V., Puiu, D., Hall, R., Kingan, S., Clavijo, B.J., Salzberg, S.L., 2017. The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. GigaScience 6, gix097.