nature genetics

Supplementary information

https://doi.org/10.1038/s41588-025-02259-2

Genomics-driven discovery of superior alleles and genes for yellow rust resistance in wheat

In the format provided by the authors and unedited

Supplementary information

Supplementary Materials

This PDF file includes:

Supplementary Materials and Methods

Supplementary Figures 1 to 9

Supplementary Materials and Methods

Wheat accessions information and genetic population

Integrated geographic and yellow rust epidemiological information identified 8 macrogeographic regions. The final panel included 2,191 accessions from the 8 macrogeographic regions: 174 from West and Central Asia, 97 from South Asia, 272 from Africa, 658 from East Asia, 106 from Latin America, 639 from Europe, 193 from North America, and 52 from Oceania. There were 1,382 (63.07%) winter, 739 (33.73%) spring, and 34 (1.55%) facultative wheat types; 684 (31.22%) were landraces, 135 (6.16%) were pre-Green Revolution before 1960, 608 (27.75%) registered 1960–2000, and 764 (34.87%) registered after 2000. The plant materials from which information was extracted are available at the Chinese Crop Germplasm Resources Information System (CGRIS), the USA National Small Grains Collection (NSGC), Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) GeneBank, International Centre for Agricultural Research in Dry Areas (ICARDA) GeneBank, and the International Maize and Wheat Improvement Center (CIMMYT) Wheat Germplasm Bank.

Genetic mapping populations recombinant inbred lines (RILs) and selected segregating RILs (heterozygous inbred families, HIFs) from five bi-parental crosses were used for validation of GWAS signals. They were Avocet S (AvS)/Baomai 6 (BM6, carrying with *Yr5b*) containing 142 ABM6-RILs and 3,645 HIFs, AvS/Xinong 3517 (XN3517, carrying with *Yr5x*) containing 162 AXN3517-RILs and 3,586 HIFs, AvS/Aikang 58 (AK58, carrying with *Yr6*) containing 128 AAK58-RILs and 4,452 HIFs, Kenong 9204 (KN9204)/Jing 411 (J411, carrying with *YrKB*) containing 188 KJ-RILs, and AvS/Flanders (FLA, carrying with *YrKB*) containing 184 AFLA-RILs and 5,426 HIFs.

Sample preparation and sequencing

Genomic DNA was extracted from seedlings of 349 wheat accessions using a Plant Genomic DNA Kit (Tiangen, GDP305) according to the manufacturer's instructions. Each sample with approximately 10 mg of DNA was used to construction a 350-bp paired-end library with the Bioruptor Pico Sonication System (diagenode). Then, the libraries were sequenced with the DNBSEQ platform of BGI-Shenzhen, generating 150-bp paired-end reads by whole genome resequencing producing ~4.9×10^{12} 100-bp paired-end reads and average sequencing coverage depth of ~11× for each accession.

Total RNA with three biological replicates was isolated using a RNeasy Plant mini kit (QIAGEN, 74904) from seedling or flag leaves of 8 accessions (parents of genetic mapping populations, and susceptible controls Mingxian 169 and Xiaoyan 22) at 48 h post inoculation with *Pst* urediniospores or water, respectively. Prior to RNA isolation, the leaf tissues from the three replicates were pooled in equal proportions. As described above, RNA-seq libraries were constructed and sequenced using BGISEQ500 on the DNBSEQ platform to generate 150-bp paired-end reads. Transcripts Per Million reads (TPM) values were calculated for each gene with Salmon (v.1.9.0)[103], with TPM less than one considered as no expression.

Genotype imputation

Genotype data from the 1,577 sequenced accessions was used as a reference panel for imputation. An integrated VCF file was created, including 614 accessions genotyped by the wheat660K SNP array and the reference panel. Beagle v.5.4 (beagle.22Jul22.46e.jar)[104] was then used to impute missing genotype calls with the following settings: 'gp=true window=40 ne=2000'. Genotype calls with probability (GP) <0.8 were considered as missing. Sites with >30% heterozygous genotype calls or >30% missing data were removed, resulting in a set of about 84,855,324 variants.

Meta-analysis

To provide a more detailed characterization and comparison of the identified loci/genes in this study, we collected information on 1,125 QTL/genes for yellow rust resistance and all cloned *R* genes or homologs for resistance to various diseases in the *Triticeae* as well as QTL/gene mapping data from more recent GWAS, QTL mapping and meta-QTL studies[105,106] from 175 publications over the past three decades. With information for individual QTL/genes, such as: (i) type and size of the mapping population, (ii)

flanking markers and their physical positions on the map, (iii) peak positions, (iv) phenotypic variance explained (PVE or R^2 value), and (v) logarithm of the odds (LOD) score (QTL mapping) or *P*-value (GWAS), we compiled an integrated reference physical map of YR resistance loci for ease of later comparison. A total of 1,125 QTL/genes and over 40 cloned resistance genes/alleles or homologous genes were positioned across the 21 wheat chromosomes based on reference genome IWGSC RefSeq v2.1 as follows: (1) the sequences of the closest linked markers or two flanking markers of the QTL confidence interval were retrieved from various databases, including WheatOmics (http://wheatomics.sdau.edu.cn), CerealsDB (https://www.cerealsdb.uk.net), GrainGenes (https://wheat.pw.usda.gov/GG3), DArT (https://www.diversityarrays.com), and URGI (https://wheat-urgi.versailles.inra.fr/); (2) physical positions of the above marker sequences obtained from the RefSeq v2.1 assembly using BLASTN; (3) physical locations of each QTL were determined by calculating the confidence intervals defined by flanking or closest linked marker positions. If the flanking interval was less than 10 Mb, the midpoint was used. Otherwise, the entire flanking interval was used to display the QTL. Presence of one independent QTL (iQTL) on the same chromosome was indicated if the QTL was in a different LD block, while QTL hotspot regions were classified if the distance was less than 10 Mb or within the Meta-QTL (MQTL) region. This distance was chosen as a conservative estimation that mirrored the scope of previous QTL mapping studies and meta-analysis studies[105,106].

Identification and prioritization of YR candidate genes

Identification of causal genes that underlie complicated agronomic traits directly from GWAS results remains difficult. Taking into consideration false-positive and false-negative issues associated with GWAS[107], we employed a comprehensive analysis pipeline using multi-omics datasets and multiple bioinformatics methods to prioritize candidate genes in QTL regions identified by GWAS. The filtering procedure was: 1) search for high-confidence genes located within the LD block and evaluate the gene expression to predict whether genes were associated with phenotypes using transcriptome datasets; 2) perform GO analysis with agriGO (v.2.0) of candidate genes and homologous genes with known function that were involved in biological pathways of plant defense responses. Enrichment significance was analyzed by Fisher's exact test. Enrichment results with more than 5 annotations and FDR <0.05 were plotted with the R package clusterProfiler (v.3.10.0); 3) using a five-level grouping method

(referred to as G1–G5) to evaluate variant effects in gene regions based on the estimated functional importance of each nucleotide polymorphism as described by Yano et al.[108]. These levels included: G1, significant MTAs in the GWAS ($-\log_{10}P \geq$the threshold value in a particular chromosome) that putatively caused amino acid conversion and alternative splicing; G2, significant MTAs in the 5′ flanking sequences (≤2 kb from the first ATG), which were considered to be promoter regions; G3, significant MTAs within the coding region but belonging to synonymous mutations, introns or 3′ noncoding sequences; G4, significant MTAs outside coding regions; and G5, polymorphic but MTA not significant; 4) calculate the degree of haplotype-based association for each potential gene using GAPIT[109].

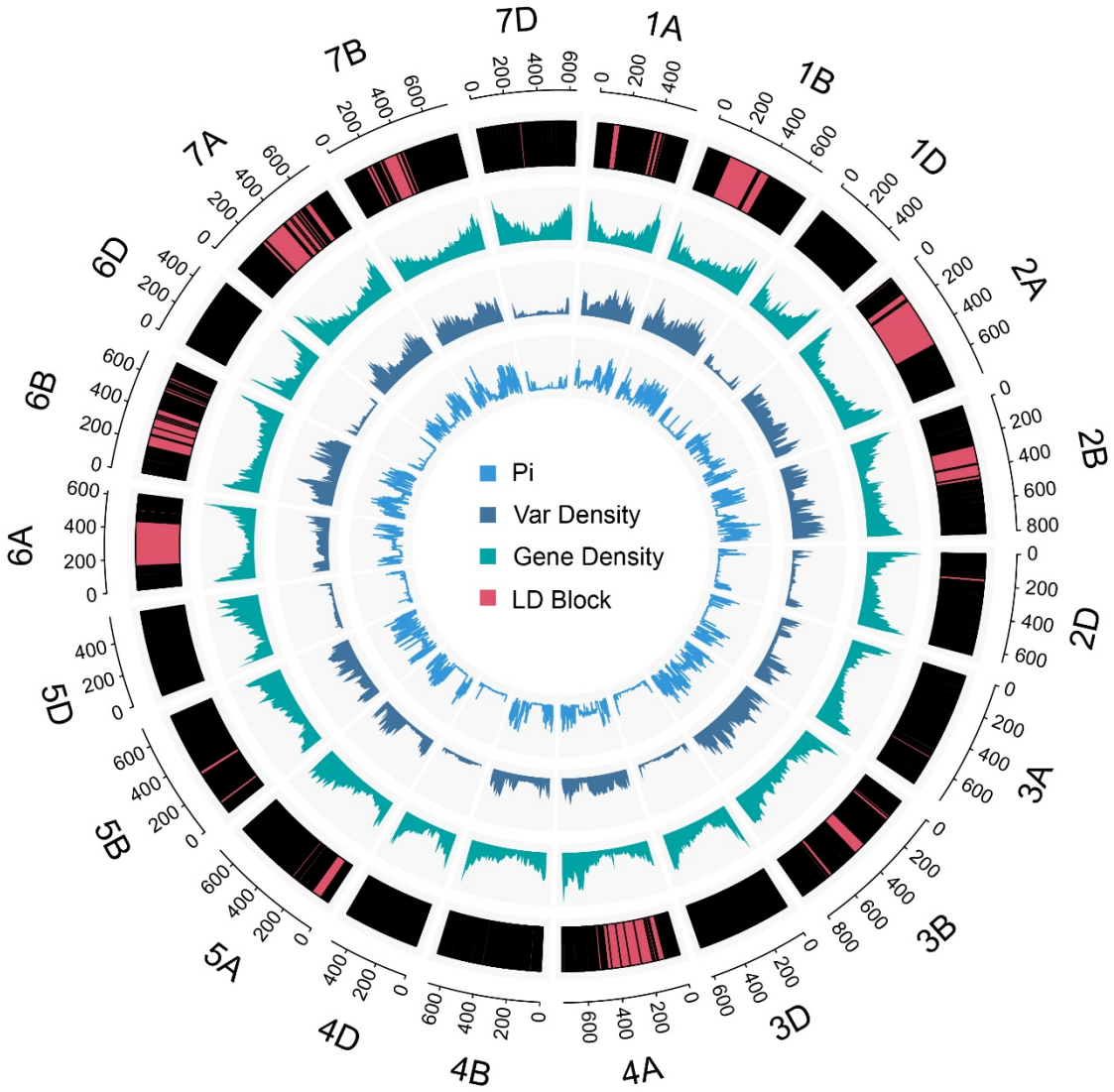
Phylogenetic analysis

The protein sequences of START and its orthologues from different plant species were extracted from the EnsemblPlants database (http://plants.ensembl.org/index.html). A phylogenetic tree was constructed using the maximum-likelihood method in the MEGA7.0 program[110] with bootstrap (1,000 replicates) and complete deletion.

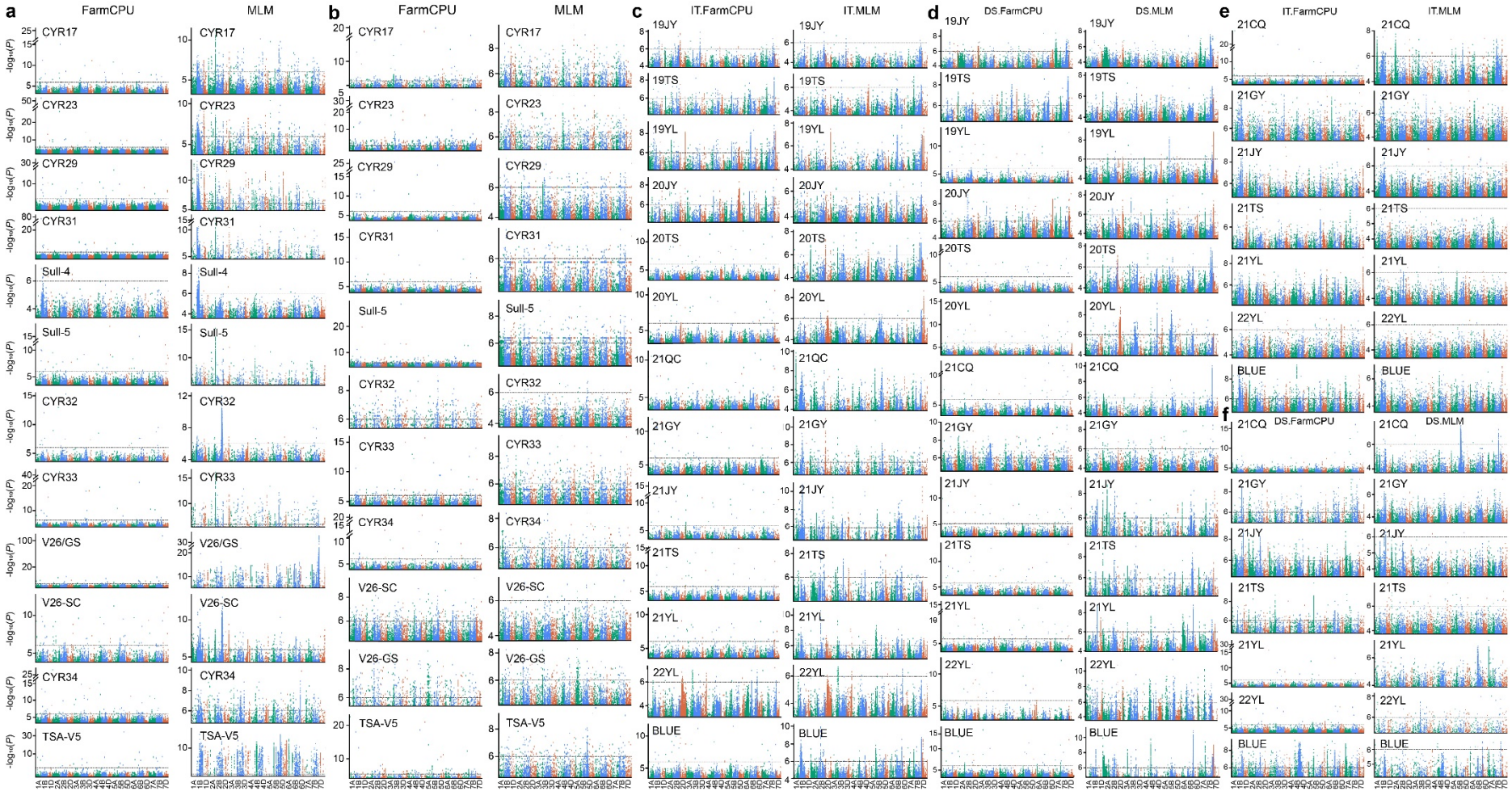
Alignments of homologs from different genomes

We collected 60 complete and scaffolded genomes from diploid, tetraploid and hexaploid wheat as well as its related species (WheatOmics, http://wheatomics.sdau.edu.cn; Supplementary Table 27)[111-114]. First, we employed the full-length target genes or additional promoter regions as reference sequence to align with each genome using BLASTN[115] and then performed multiple sequence alignments using Muscle[116] before presenting the results visually to deliver a detailed analysis of the evolutionary conservation and variability of homologous genes across different genomes. Throughout this process, we checked for the presence of the start “ATG” codon and stop codons to ensure completeness of the gene sequences. Finally, we determined the classification of the homologous gene sequences based on similarity clustering in the multiple sequence alignments.

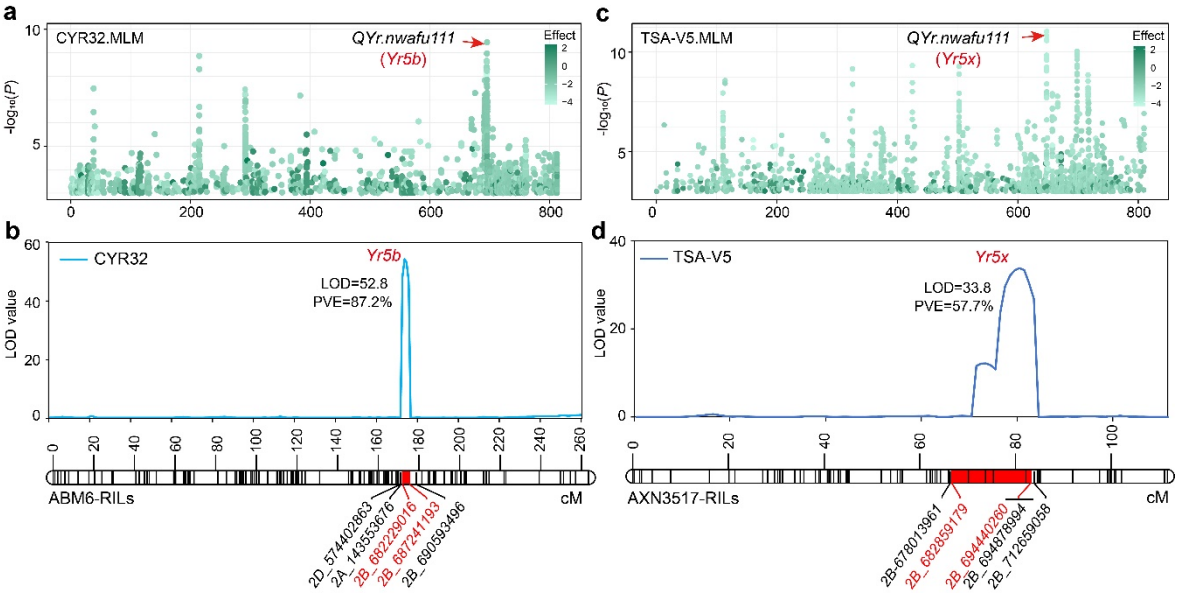
Supplementary Figures



Supplementary Fig. 1 Genetic diversity in 2,191 hexaploid wheat genomes. Circos diagram from the outer ring to the inner ring showing physical chromosome map, genome-wide physical distributions of LD blocks, gene density, genome-wide physical distributions of filtered variants, and average nucleotide diversity in each chromosome.

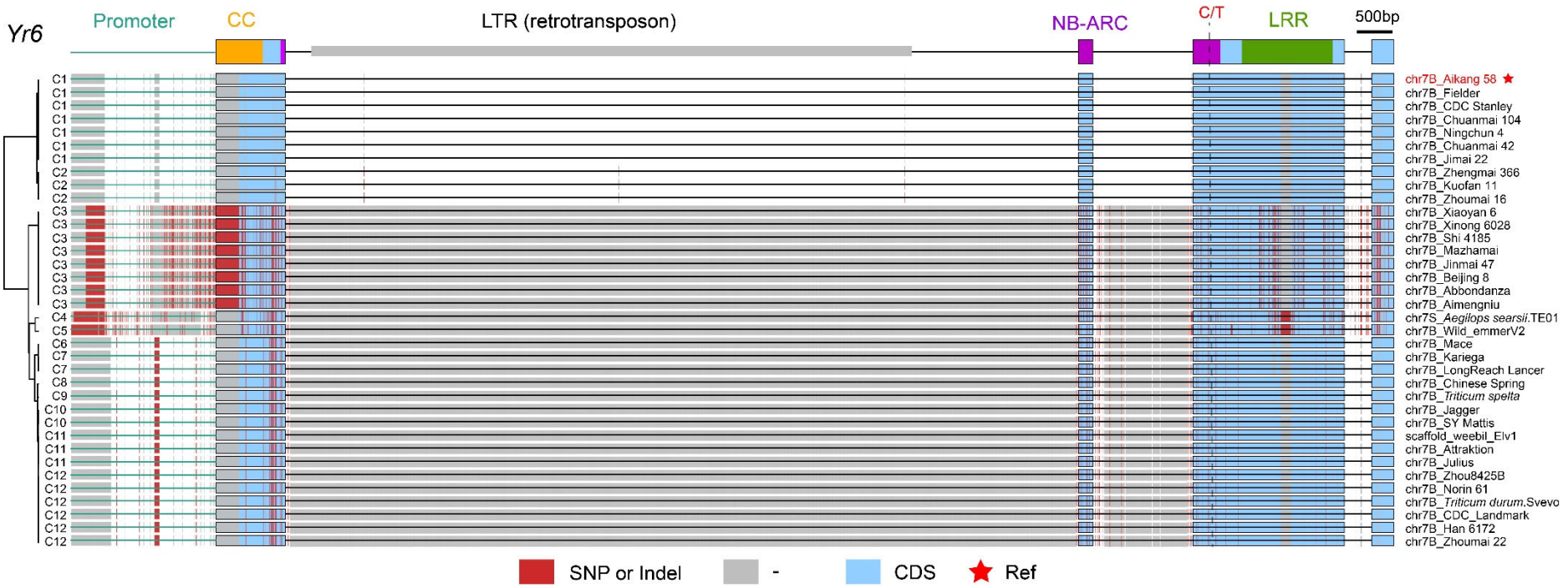

a
FarmCPU
MLM
CYR17
CYR23
CYR29
CYR31
Sull-4
Sull-5
CYR32
CYR33
V26/GS
V26-SC
CYR34
TSA-V5
b
FarmCPU
MLM
CYR17
CYR23
CYR29
CYR31
Sull-5
CYR32
CYR33
CYR34
V26-SC
V26-GS
TSA-V5
c
IT.FarmCPU
IT.MLM
19JY
19TS
19YL
20JY
20TS
20YL
21QC
21GY
21JY
21TS
21YL
22YL
BLUE
d
DS.FarmCPU
DS.MLM
19JY
19TS
19YL
20JY
20TS
20YL
21CQ
21GY
21JY
21TS
21YL
22YL
BLUE
e
IT.FarmCPU
IT.MLM
21CQ
21GY
21JY
21TS
21YL
22YL
BLUE
f
21CQ
DS.FarmCPU
21CQ
DS.MLM
21GY
21JY
21TS
21YL
22YL
BLUE

Supplementary Fig. 2 Genome-wide association study (GWAS) of yellow rust resistance loci detected with different *Pst* races, multiple environments and different association mapping panels using FarmCPU and MLM models. Manhattan plots of GWAS results for: **a**, Seedling resistance using 12 *Pst* races. **b**, Seedling resistance of landraces to 11 *Pst* races. **c**, IT across individual environments (2019-Jiangyou (19JY), 2019-Tianshui (19TS), 2019-Yanlging (19YL), 2020-Jiangyou (20JY), 2020-Tianshui (20TS), 2020-Yanlging (20YL), 2021-Chongqing (21CQ), 2021-Guiyang (21GY), 2021-Jiangyou (21JY), 2021-Tianshui (21TS), 2021-Yangling (21YL), 2022-Yangling (22YL)) and best linear unbiased estimator (BLUE). **d**, DS across the same environments, and BLUE. **e**, IT of landraces across 5 environments (21CQ, 21GZ, 21JY, 21TS, 21YL) and BLUE. **f**, DS of landraces across the same environments, and BLUE.



Supplementary Fig. 3 GWAS and QTL mapping of *Yr5b* and *Yr5x*.

a, Genomic regions in chromosome 2B associated with resistance to *Pst* race CYR32. *QYr.nwafu111* (*Yr5b*) is indicated by red arrow. **b**, QTL mapping of *Yr5b* in the Avocet S/Baomai 6 (ABM6) RIL population. The candidate region is shown in red. **c**, Genomic regions in chromosome 2B associated with resistance to *Pst* race TSA-V5. *QYr.nwafu111* (*Yr5x*) is indicated by red arrow. **d**, QTL mapping of *Yr5x* in the Avocet S/Xinong 3517 (AXN3517) RIL population. The candidate region is shown in red.

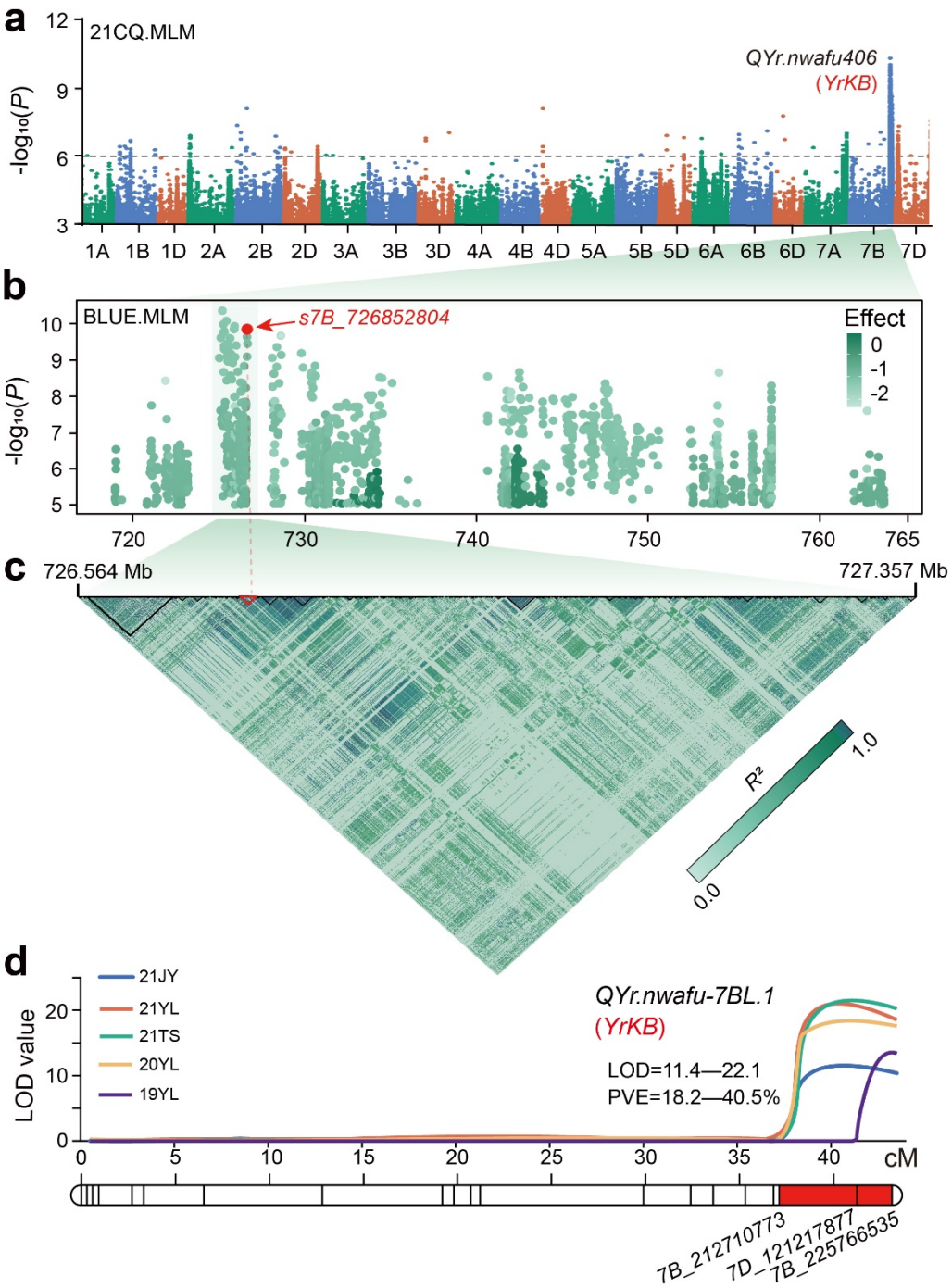


Supplementary Fig. 4 Multiple alignment *Yr5x* homologs across complete and scaffolded genomes.

Twenty-two homologous sequences from different genomes were identified including *Yr5a*, *Yr5b* and *Yr5x*. The stop mutation nucleotide is labeled. Lightskyblue boxes, coding DNA sequence (CDS); lines, introns; silver boxes, gaps; red lines, SNPs and Indels; other colored boxes in the gene structure, different domains; red five-pointed star, reference sequence.

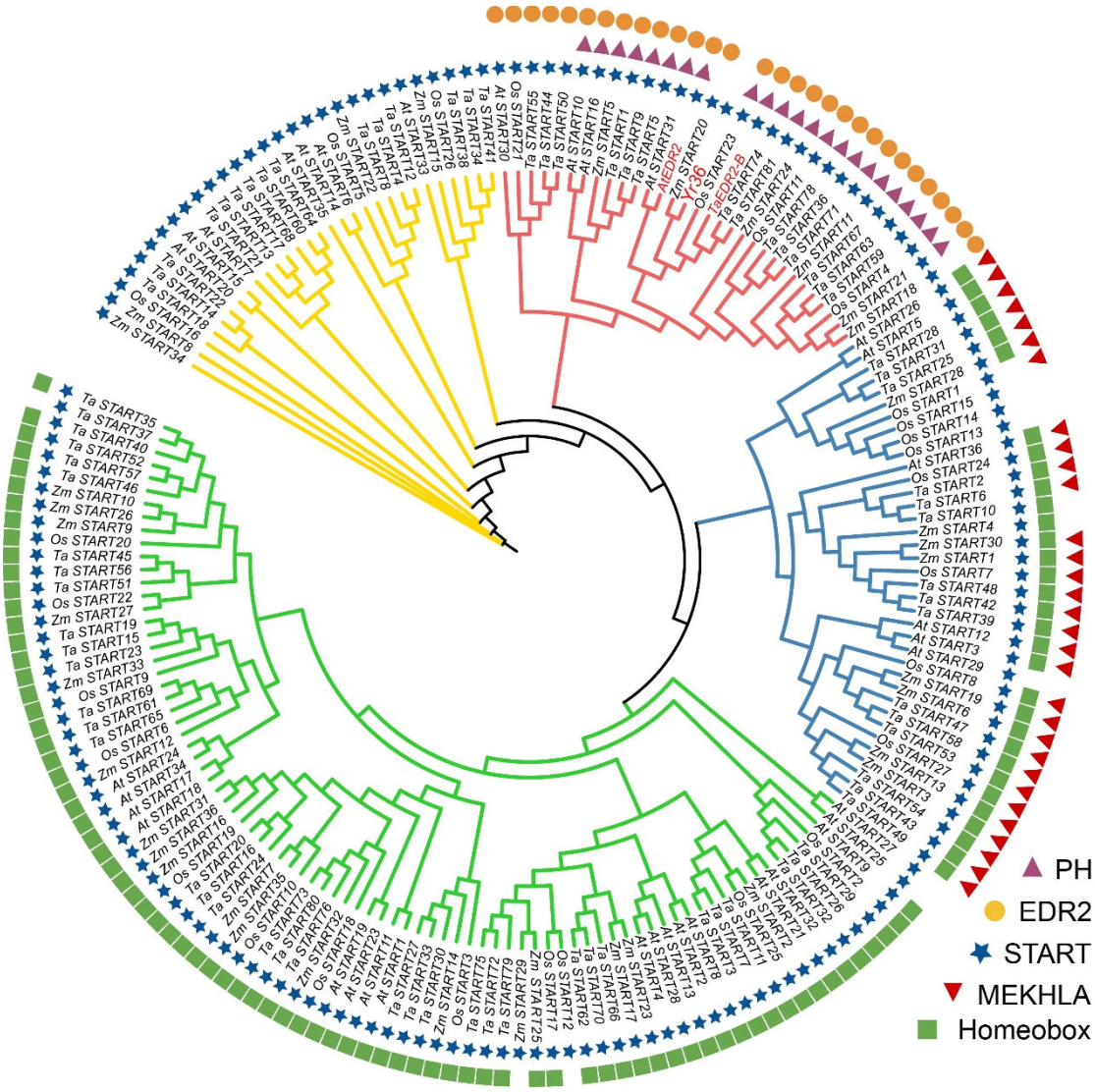
Supplementary Fig. 5 Multiple alignment of *Yr6* homologs across different complete and scaffolded genomes.

Thirty-seven homologous sequences of *Yr6* in different genomes. The C/T nucleotide significantly associated with YR resistance is labeled. Lightskyblue boxes, coding DNA sequence (CDS); lines, introns; silver boxes, gaps; red lines, SNPs and Indels; other colored boxes in the gene structure, different domains; red five-pointed star, reference sequence.



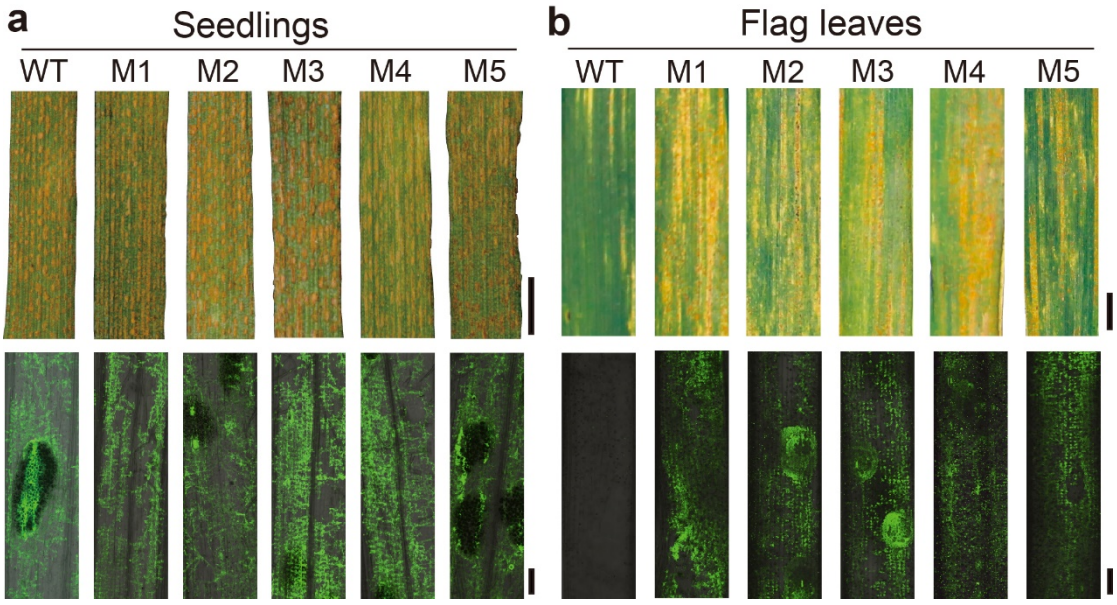
Supplementary Fig. 6 GWAS and QTL mapping of *YrKB*.

a, GWAS signals for *QYr.nwafu406* (*YrKB*) detected using 21CQ_BLUE. **b**, Regional Manhattan plot of the *YrKB* genomic region on chromosome 7B. **c**, Pairwise LD analysis. The significant variant is indicated by a red arrow. **d**, QTL mapping of *YrKB* using an Avocet S/Flanders (AFLA) RIL population. The candidate region is colored red.



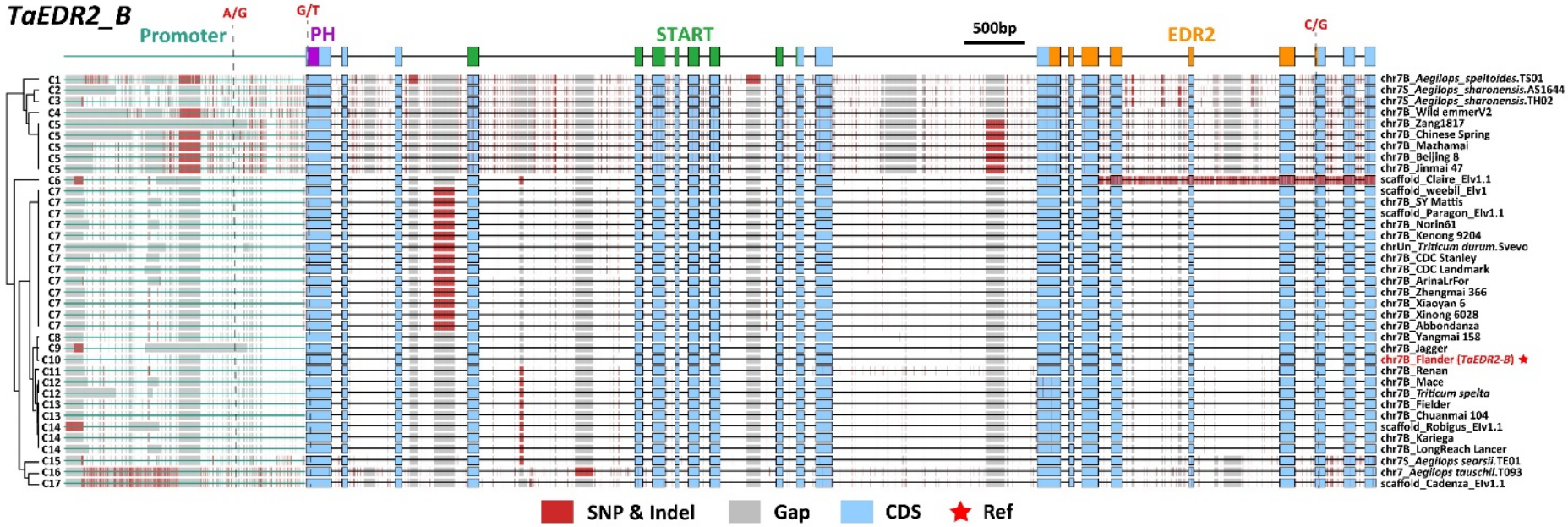
Supplementary Fig. 7 Phylogenetic analysis of START proteins from different plant species.

TaEDR2-B and *Yr36* (*TaWKS1*) are indicated in red script. Purple triangles: PH domain; orange circles: EDR2 domain; blue stars: START domain; cyan inverted triangles: MEKHLA domain; green rectangles: homeobox domain.



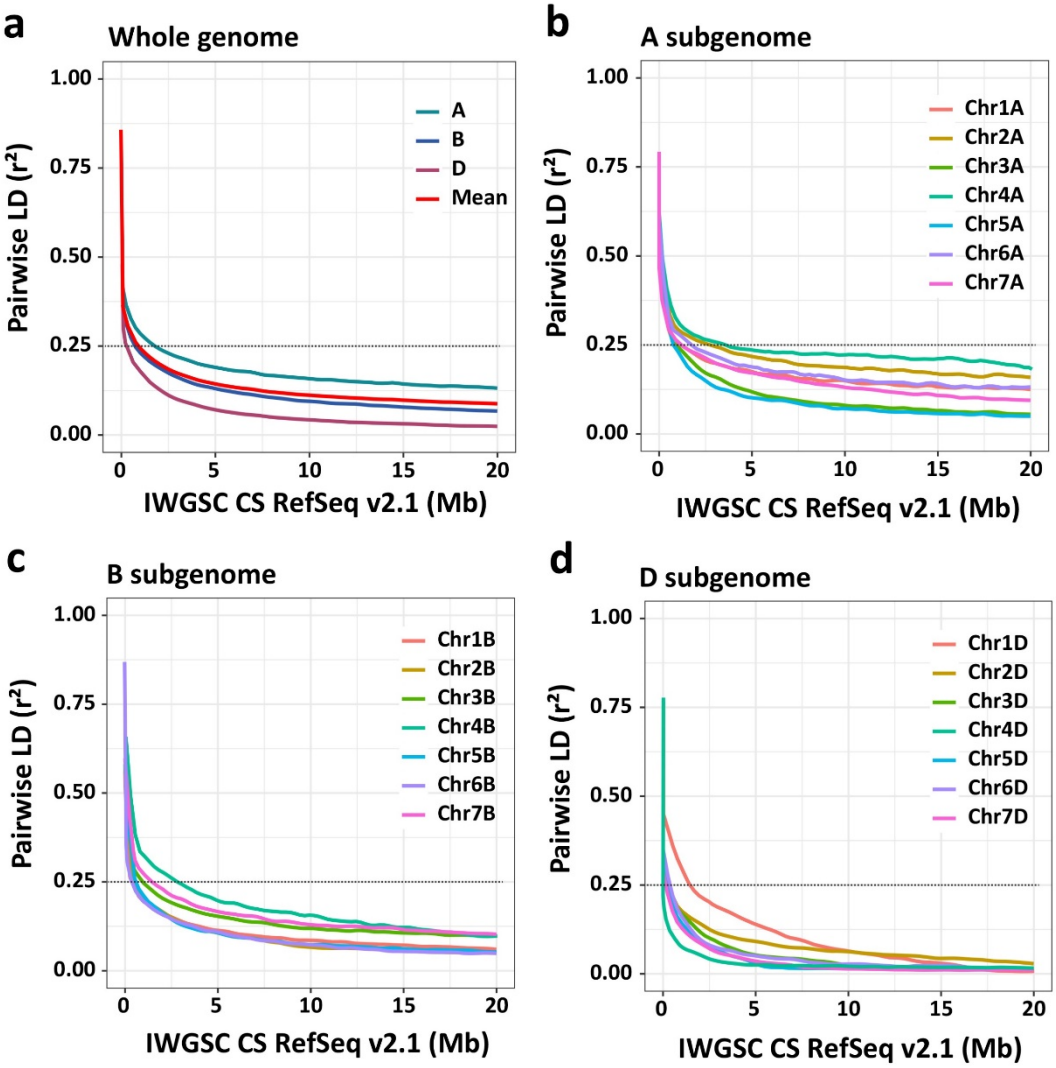
Supplementary Fig. 8 Functional validation of *TaEDR2-B* using ethyl methane sulfonate (EMS) mutants.

a, Seedling responses of Jing 411 (WT) and five independent EMS mutants to mixed *Pst* races at 16 dpi (Scale bar, 1 cm) and corresponding histological observations conducted using WGA staining (Scale bar, 100 μm). **b**, Adult plant responses (Scale bar, 1 cm) and WGA staining (Scale bar, 100 μm) of the same genotypes showing less hyphal development.



Supplementary Fig. 9 Multiple alignment of *TaEDR2-B* homologs across complete and scaffolded genomes.

Thirty-nine homologous sequences from different genomes were identified including *TaEDR2-B.* The nucleotides significantly associated with YR resistance are labeled. Lightskyblue boxes, coding DNA sequence (CDS); lines, introns; silver boxes, gaps; red lines, SNPs and Indels; other colored boxes in the gene structure, different domains; red five-pointed star, reference sequence ‘Flanders’.



Supplementary Fig. 10 Genome-wide average linkage disequilibrium (LD) decay over physical distances

a, Whole genome and three subgenomes. **b**, A subgenome. **c**, B subgenome. **d**, D subgenome.

Additional References (Supplementary Materials and Methods and Supplementary Figures)

103. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419 (2017).

104. Browning, B.L., Tian, X., Zhou, Y. & Browning, S.R. Fast two-stage phasing of large-scale sequence data. *Am J Hum Genet* **108**, 1880-1890 (2021).

105. Pal, N. et al. Meta-QTLs for multiple disease resistance involving three rusts in common wheat (*Triticum aestivum* L.). *Theor Appl Genet* **135**, 2385-2405 (2022).

106. Kumar, S. et al. Comprehensive meta-QTL analysis for dissecting the genetic architecture of stripe rust resistance in bread wheat. *BMC Genomics* **24**, 1-19 (2023).

107. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res* **20**, 393-402 (2010).

108. Yano, K. et al. Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nat Genet* **48**, 927-34 (2016).

109. Wang, J. & Zhang, Z. GAPIT version 3: Boosting power and accuracy for genomic association and prediction. *Genom Proteom Bioinf* **19**, 629-640 (2021).

110. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* **33**, 1870-1874 (2016).

111. Jiao, C. et al. Pangenome bridges wheat structural variations with habitat and breeding. *Nature* **637**, 384-393 (2025).

112. Li, G. et al. Genomic analysis of Zhou8425B, a key founder parent, reveals its genetic contributions to elite agronomic traits in wheat breeding. *Plant Commun* **6**, 1-19 (2025).

113. Liu, Z. et al. Chromosome-level assembly of the synthetic hexaploid wheat-derived cultivar Chuanmai 104. *Sci Data* **11**, 1-11 (2024).

114. Walkowiak, S. et al. Multiple wheat genomes reveal global variation in modern breeding. *Nature* **588**, 277-283 (2020).

115. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 1-9 (2009).

116. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797 (2004).