nature research

https://doi.org/10.1038/s41588-020-00722-w

Supplementary information

Triticum population sequencing provides insights into wheat adaptation

In the format provided by the authors and unedited

Supplementary Information

Title

Triticum population sequencing provides insights into wheat adaptation

Author List

Yao Zhou[1,7,#], Xuebo Zhao[1,2,#], Yiwen Li[1], Jun Xu[1,2], Aoyue Bi[1,2], Lipeng Kang[1,2], Daxing Xu[1,2], Haofeng Chen[3,8], YingWang[2], Yuan-ge Wang[3], Sanyang Liu[4], Chengzhi Jiao[5], Hongfeng Lu[5], Jing Wang[1], Changbin Yin[1], Yuling Jiao[2,3,*], Fei Lu[1,2,6,*]

Institutions

[1]State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Innovative Academy of Seed Design, Chinese Academy of Sciences, Beijing.

[2]University of Chinese Academy of Sciences, Beijing.

[3]State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Innovative Academy of Seed Design, Chinese Academy of Sciences, Beijing

[4]Annoroad Gene Technology Corporation, Beijing.

[5]Novogene Bioinformatics Institute, Beijing.

[6]CAS-JIC Centre of Excellence for Plant and Microbial Science (CEPAMS), Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing.

Present addresses

[7]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.

[8]National Research Institute for Family Planning, Beijing.

Equal contributions statement

[#]These authors contributed equally to this work.

Corresponding author statement

Fei Lu (flu@genetics.ac.cn) and Yuling Jiao (yljiao@genetics.ac.cn) are corresponding authors.

INVENTORY

Supplemental Information contains the Supplemental Data (including 10 Extended Data Figures, 6 Supplementary Notes, and 14 Supplementary Figures).

Extended Data Figure 1 and 2 are related to Supplementary Note 1.
Extended Data Figure 3 is related to Figure 2.
Extended Data Figure 4 is related to Figure 3.
Extended Data Figure 5-7 are related to Figure 4.
Extended Data Figure 8 and 9 are related to Figure 5.
Extended Data Figure 10 is related to Figure 7.

Supplementary Notes 1-6

Supplementary Figure 1 is related to Figure 1.
Supplementary Figure 2 is related to Supplementary Note 1.
Supplementary Figure 3 and 4 are related to Figure 2.
Supplementary Figure 5 is related to Figure 3.
Supplementary Figure 6-10 are related to Figure 4.
Supplementary Figure 11 and 12 are related to Figure 7.
Supplementary Figure 13 is related to Supplementary Note 5.
Supplementary Figure 14 is related to Figure 8.

Supplementary Note 1

A comprehensive whole-genome genetic variation map of wheat. To obtain genome-wide genetic variations and fully present adaptive signatures of bread wheat, we constructed a comprehensive whole-genome genetic variation map of wheat (VMap 1.0) by whole-genome sequencing of all 414 accessions with an average coverage at 6x (Supplementary Table 3). A custom-designed pipeline for cross-ploidy variation discovery was implemented to overcome the multiple ploidy levels within the collection and ensure high accuracy of variation calling (Extended Data Fig. 1). First, to obtain a unified coordinate system in VMap 1.0, genetic variations were discovered from each genome/subgenome group by taking corresponding genome/subgenomes from the wheat genome IWGSC RefSeq v1.0[1] as its own reference. For example, reads from AA taxa were aligned to the A subgenome of IWGSC RefSeq v1.0 for variation discovery. Next, to avoid reference bias of variation calling where species/samples dissimilar to the reference genome generally have more unreliable read mapping and spurious variation calls, only genetic variations in syntenic regions within each lineage were retained (Extended Data Fig. 2a,b). Then, variations from each group were merged into a variation library to construct VMap 1.0 by scanning through all bam files of 414 accessions. Respectively, about 13%, 44%, and 48% of genomic sites were defined as syntenic in A, B, and D subgenome, which helped to identify a total of ~104M single nucleotide polymorphisms (SNPs). The relative low proportion of syntenic sites in the A subgenome was due to the long genetic distance between AA taxa and the reference genome (Fig. 2a). In the final data set, the A, B, and D subgenomes contained ~30M, ~49M, and ~25M SNPs, respectively (Supplementary Fig. 2). The large difference of SNP number between subgenomes partially reflected the asymmetric genetic diversity in the subgenomes of bread wheat[2]. The false positive error rate of variant calling, i.e., the proportion of segregating sites in the reference accession, Chinese Spring, was 0.024%, which was similar to the error rate of SNP calling in cassava[3]. The low error rate was also verified by the lowest genetic divergence, i.e., identity by state (IBS) distance, of 0.0037 between the sample of Chinese Spring and the reference genome compared with all other genotyped accessions (Extended Data Fig. 2c).

Supplementary Note 2

Evaluation of ABBA-BABA statistics to detect gene flow. Two major ABBA-BABA models to identify genomic regions of introgression were evaluated, however, consistent with earlier assessments[4], the *D* statistic[5] showed inflated value for regions of low-diversity (Supplementary Fig. 6). Hence, the more stable statistic, f_d [4], which signifies gene flow when $0 < f_d < 1$, was used to calculate the fraction of introgression in landraces in 100-SNP windows.

Supplementary Note 3

Quantifying the magnitude of introgression at individual-level. By testing introgression of each landrace from each wild individual, the f_d values across 100-SNP windows showed an explicit binary distribution, with two extremes representing non-introgression and introgression segments, respectively (Supplementary Fig. 10). Introgression window in a landrace could be identified if at least one wild individual showed $f_d = 1$ at a specific window.

Supplementary Note 4

Archaeological records and historical prevalence of introgression donor populations. The earliest remains of hexaploid wheat (8,800 ~ 8,400 BP) were discovered from Can Hassan III in southern Turkey and Abu Hureyra in Syria[6,7], which overlapped with the habitats of wild emmer in the Fertile Crescent. Later records of hexaploid wheat were unearthed from Southwest of the Caspian Sea, where strangulata is distributed. Domesticated emmer was abundant in the Middle East during the Pre-Pottery Neolithic B period (10,800 ~ 8,500 BP) and maintained as a major crop for a few thousand years, until it was largely replaced by durum (about 3,000 BP), a cultivated subspecies of free-threshing tetraploids[6]. These archaeological evidences reflect the historical contact between bread wheat and these wild populations.

Supplementary Note 5

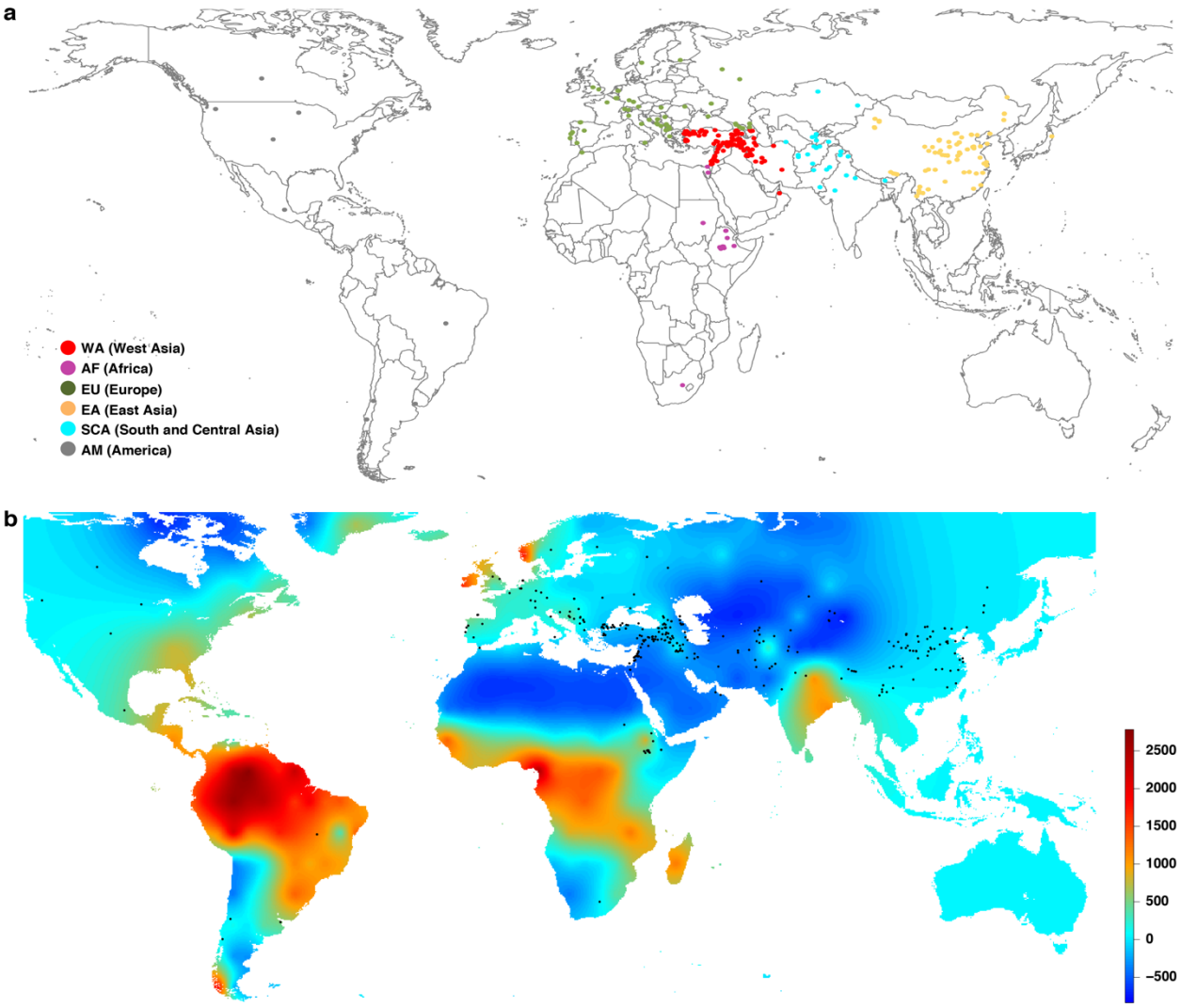
The biological function of genes showing adaptive convergence. The jointly selected genes during domestication and improvement were expected to be very important to agronomic traits of wheats. Gene ontology (GO) analysis showed that the 312 jointly selected genes of the paired domestication events were enriched for cell wall functions and glucan metabolic processes (Supplementary Fig. 13a), which could be explained by the rise of erect growth habit during early vegetative growth phase in domesticated wheats and thus requiring the strengthened stalk[8]. Meanwhile, the 473 jointly selected genes of the paired improvement events showed enrichment mostly in stimulus-response and metabolic process (Supplementary Fig. 13b), possibly reflecting recent breeding intensification efforts for disease resistance and yield.

Supplementary Note 6

Convergent evolution of brittle rachis trait. We found that the three loss-of-function alleles dominated domesticated einkorn, domesticated emmer, and its descendant, hexaploid wheats (Fig. 8b,c). However, as a species with typical non-BR phenotype, bread wheat did not show any loss-of-function alleles in *TtBtr1-D*, the ortholog of *Btr1* in the D subgenome (Fig. 8a,d). As *Btr1* is expressed in spike[9], we conducted a gene expression analysis of *TtBtr1-A*, *TtBtr1-B*, and *TtBtr1-D* in spike of bread wheat. It turned out that the dysfunction of *TtBtr1-D* was not required for the non-BR phenotype in bread wheat (Fig. 8e), which was confirmed by the non-BR phenotype of a synthetic hexaploid wheat developed from hybridization of domesticated emmer and *Ae. tauschii* (Supplementary Fig. 14).

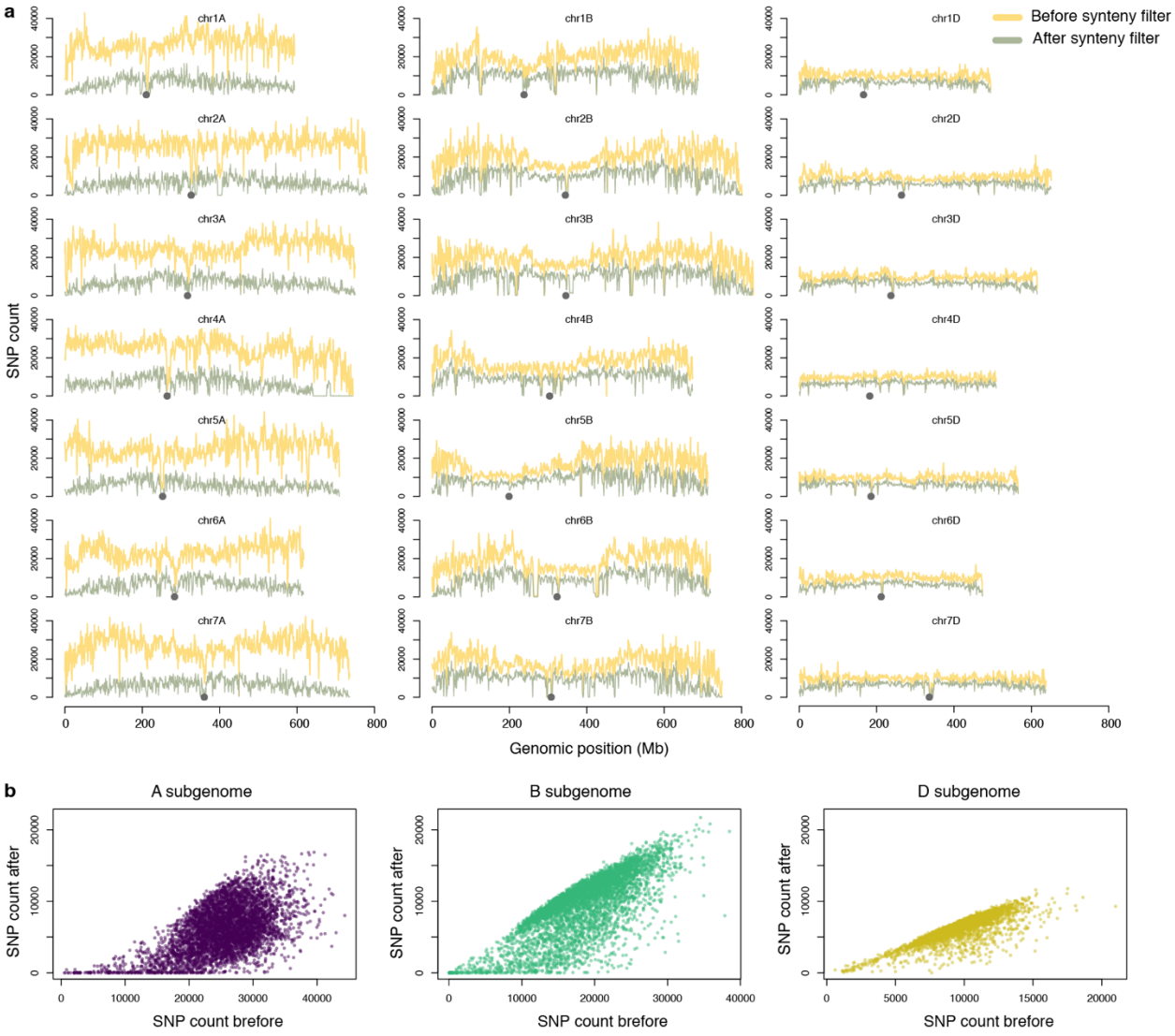
Selective sweeps identified from the domestication of emmer and einkorn exhibited strong signals of selection around *TtBtr1-A*, *TtBtr1-B,* and *TmBtr1-A* (Fig. 8f,g,h). Among them, *TtBtr1-A* and *TmBtr1-A* were in the list of jointly selected genes between the paired domestication events (Fig. 7e). In contrast, we did not find a selective sweep around the gene *TrBtr1-D* (Fig. 8i). Hence, the parallel fixation of loss-of-function alleles of *Btr1* orthologs in domesticated einkorn, domesticated emmer, and bread wheat demonstrate strong convergent adaptation of wheats to breeding selection.

Supplementary Figure 1



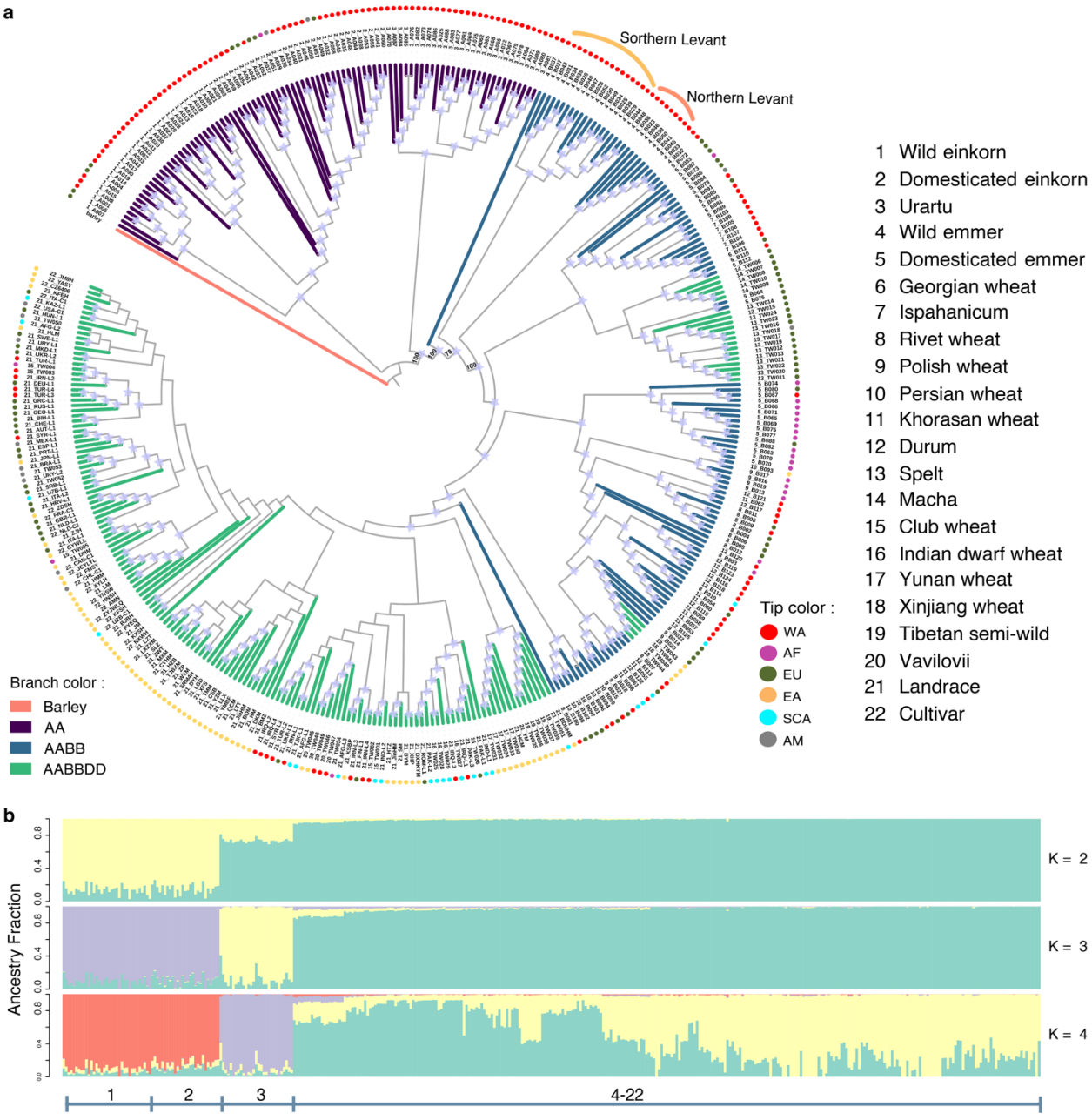
Supplementary Fig. 1. The worldwide collection of wheat accessions. **a**, The geographic distribution of all accessions. Dots with red, purple, green, brown, cyan, grey colors represent accessions from West Asia (WA), Africa (AF), Europe (EU), East Asia (EA), South, and Central Asia (SCA), and America (AM). **b**, The geographic projection of environmental factors. A total of 19 environmental factors from WorldClim (https://www.worldclim.org/) were collected every two degrees of latitude and longitude. The colors present the values of the first principle component of these environmental factors. Black dots represent all accessions. Maps were created using R packages 'OpenStreetMap' and 'rworldmap'.

Supplementary Figure 2



Supplementary Fig. 2. The effect of synteny filer on SNPs density. **a**, The whole genome distribution of SNPs density. The yellow and grey lines represent the distribution of SNPs density before and after the application of synteny filter, respectively. **b**, The correlation between SNPs density before and after the application of the synteny filter for each subgenome. The SNPs density was calculated in 1-Mb windows along the chromosomes.

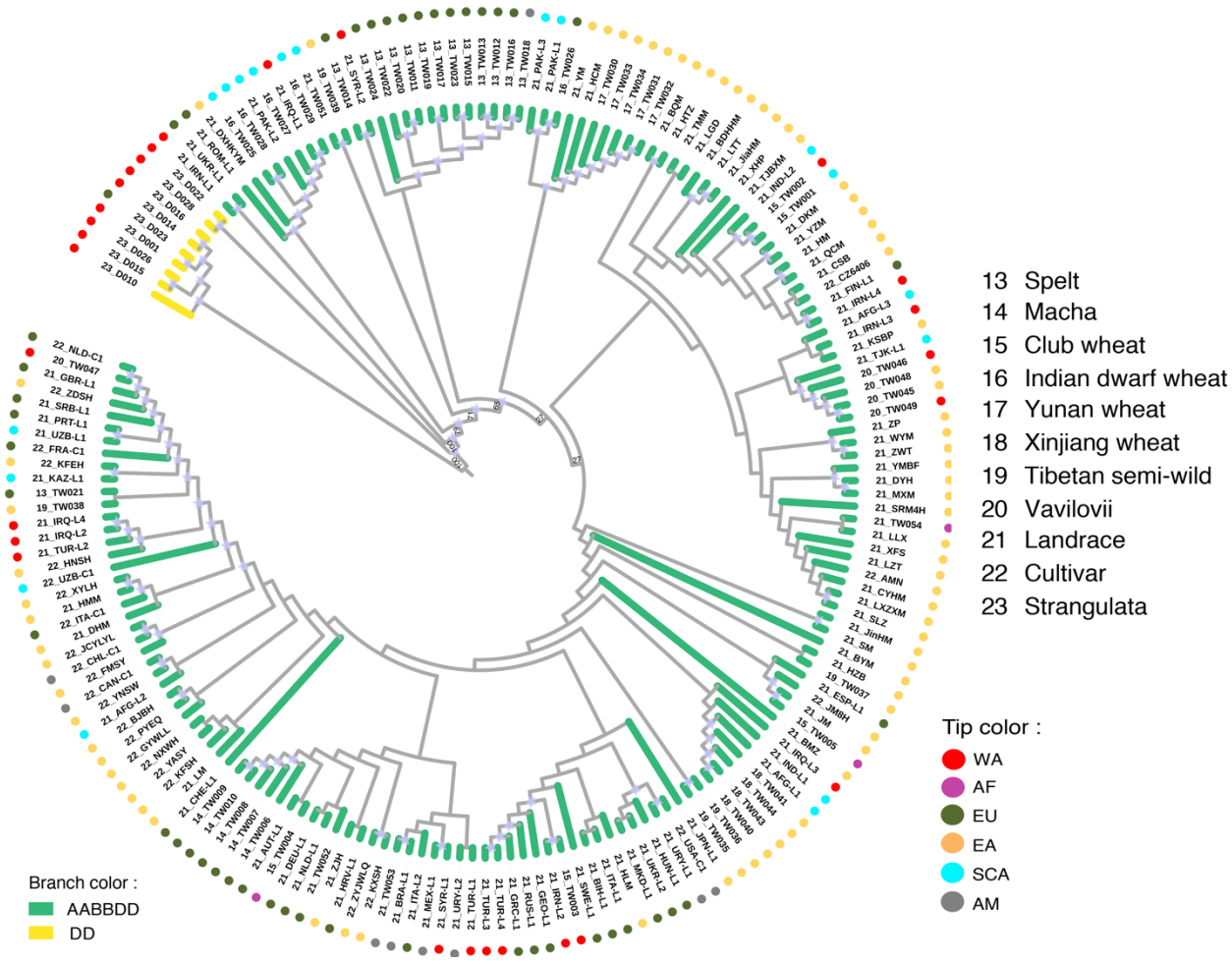
Supplementary Figure 3



Supplementary Fig. 3. Population structure of A lineage. **a**, Phylogenetic tree of A lineage. Barley was used as the outgroup. Species/subspecies were represented by numbers from 1 to 22. Tip colors of the phylogeny represent sampling region of each accession, while branch colors represent ploidy levels. Branches with reliable bootstrap value (> 50) are labeled with a purple pentagram at corresponding nodes. WA: West Asia; AF: Africa; EU: Europe; EA: East Asia; SCA: South and Central Asia; AM: America. **b,**

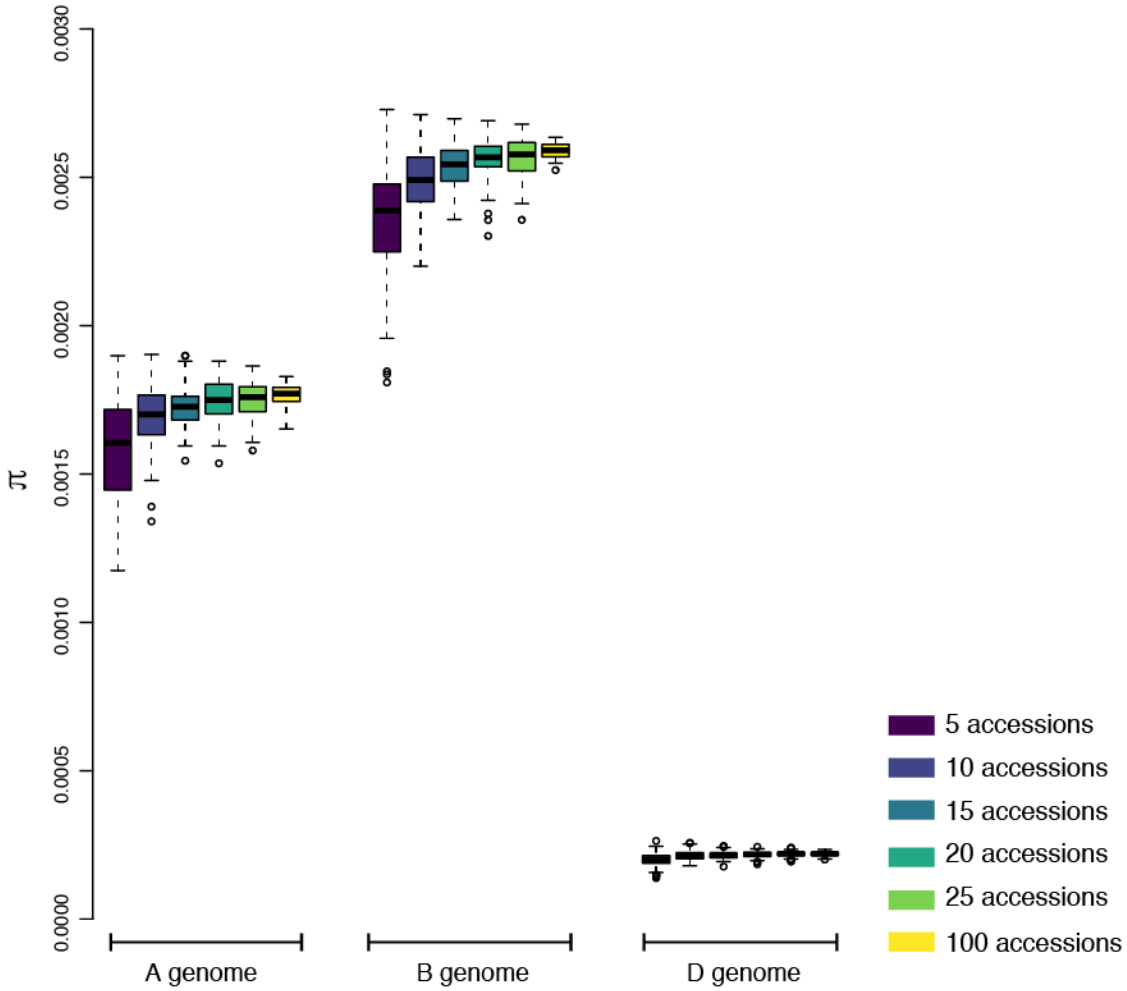
Ancestry coefficient analysis of A lineage with K = 4. Species are labeled with their corresponding numbers and colors.

Supplementary Figure 4



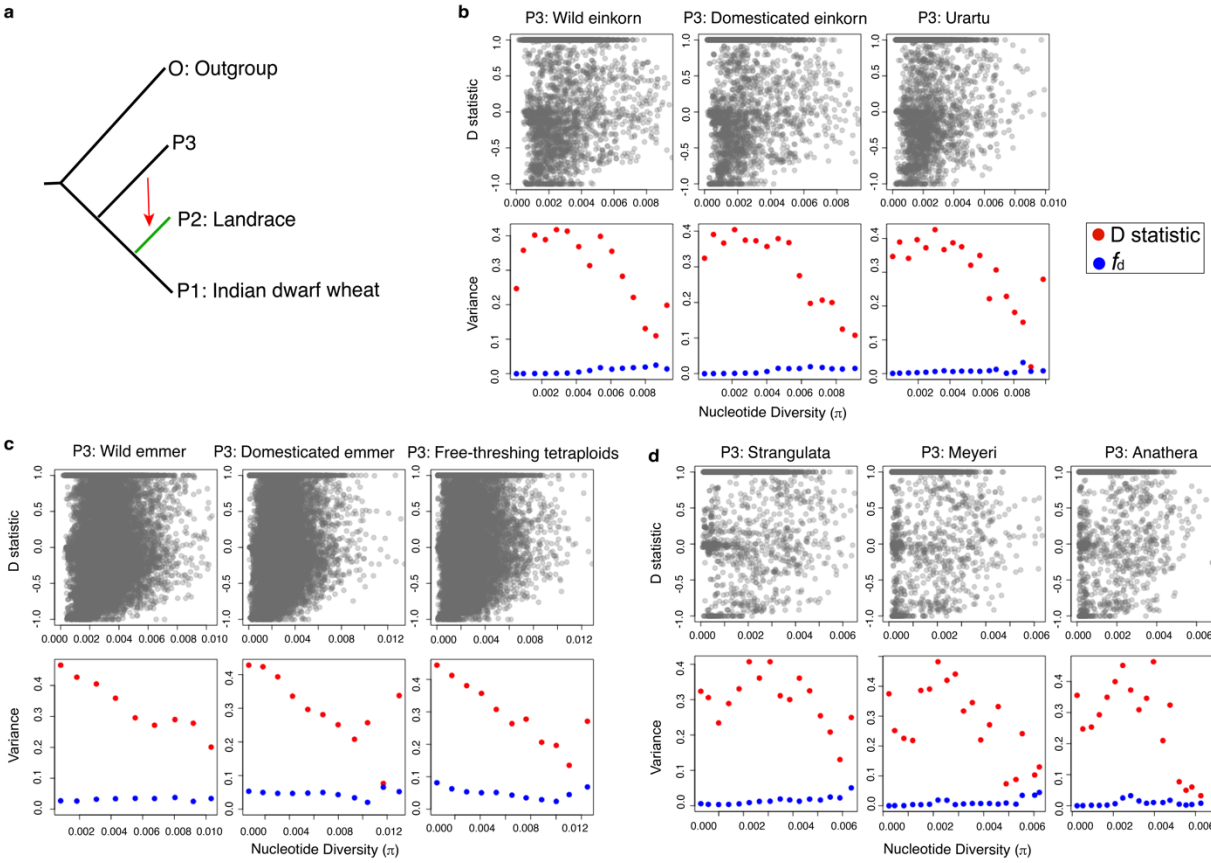
Supplementary Fig. 4. Phylogenetic tree of D lineage. Strangulata was used as the outgroup of the phylogeny of D lineage. Species/subspecies are represented by numbers from 1 to 25. Tip colors of the phylogeny represent sampling region of each accession (WA: West Asia; AF: Africa; EU: Europe; EA: East Asia; SCA: South and Central Asia; AM: America), while branch colors represent ploidy levels. Branches with reliable bootstrap value (> 50) are labeled with a purple pentagram at corresponding nodes.

Supplementary Figure 5



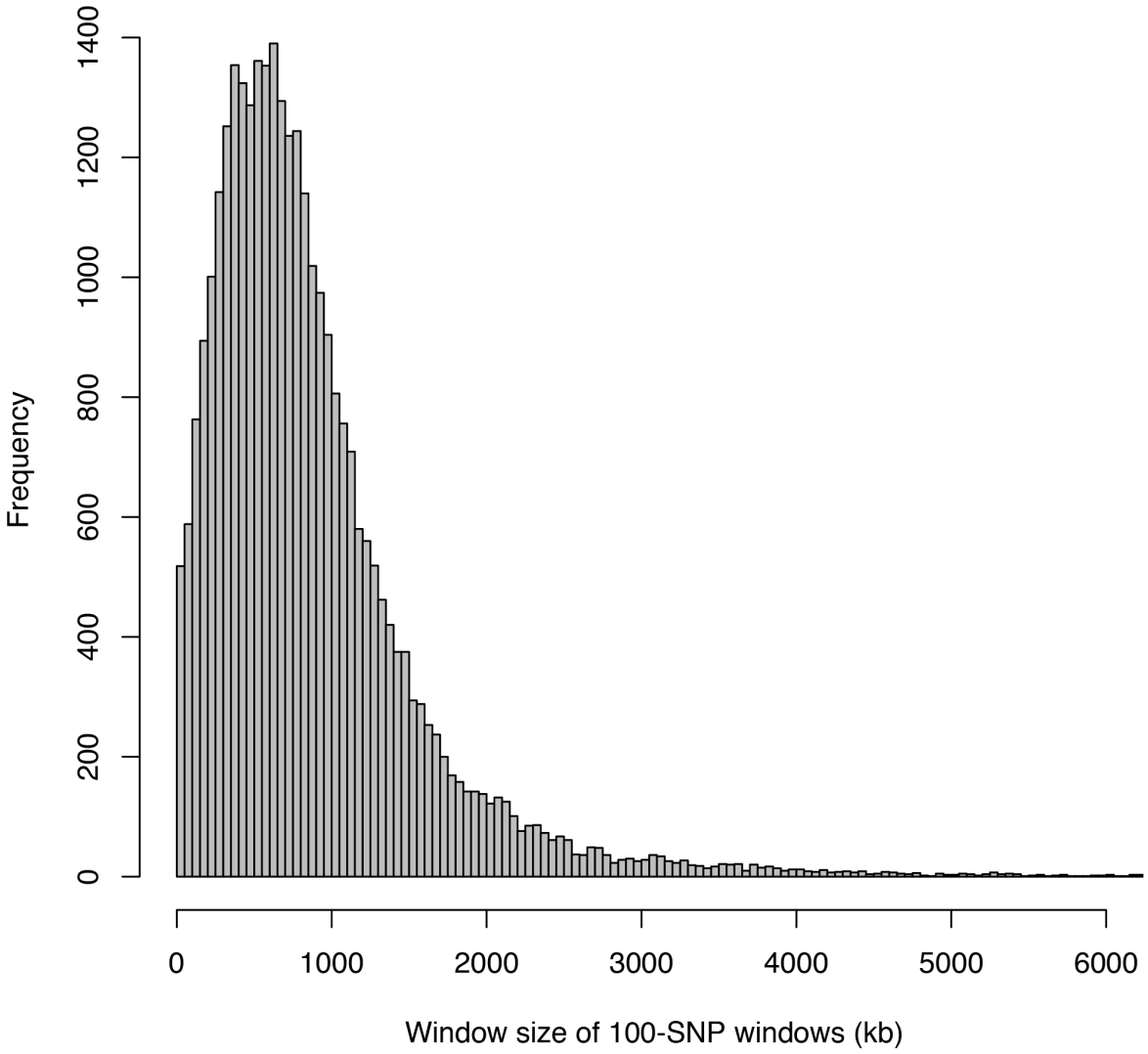
Supplementary Fig. 5. The effect of sample size on nucleotide diversity estimation. Subpopulations with a sample size of 5, 10, 15, 20, 25, and 100 were randomly sampled from bread wheat for 100 times, respectively. They were used to calculate the mean and variance of nucleotide diversity (π). The boxplots represent the π estimates in each subgenome.

Supplementary Figure 6



Supplementary Fig. 6. Bias of *D* statistic in low-diversity genomic regions. **a,** Four-taxon topology for estimating *D* statistic and f_d with Indian dwarf wheat as P1 and European (EU) landraces as P2. **b,** *D* statistic (top), variance of *D* statistic, and f_d (bottom) as a function of nucleotide diversity in A lineage. **c,** *D* statistic (top), variance of *D* statistic, and f_d (bottom) as a function of nucleotide diversity in AB lineage. **d,** *D* statistic (top), variance of *D* statistic, and f_d (bottom) as a function of nucleotide diversity in D lineage. Red dots represent variance of *D* statistic and blue dots represent variance of f_d.

Supplementary Figure 7

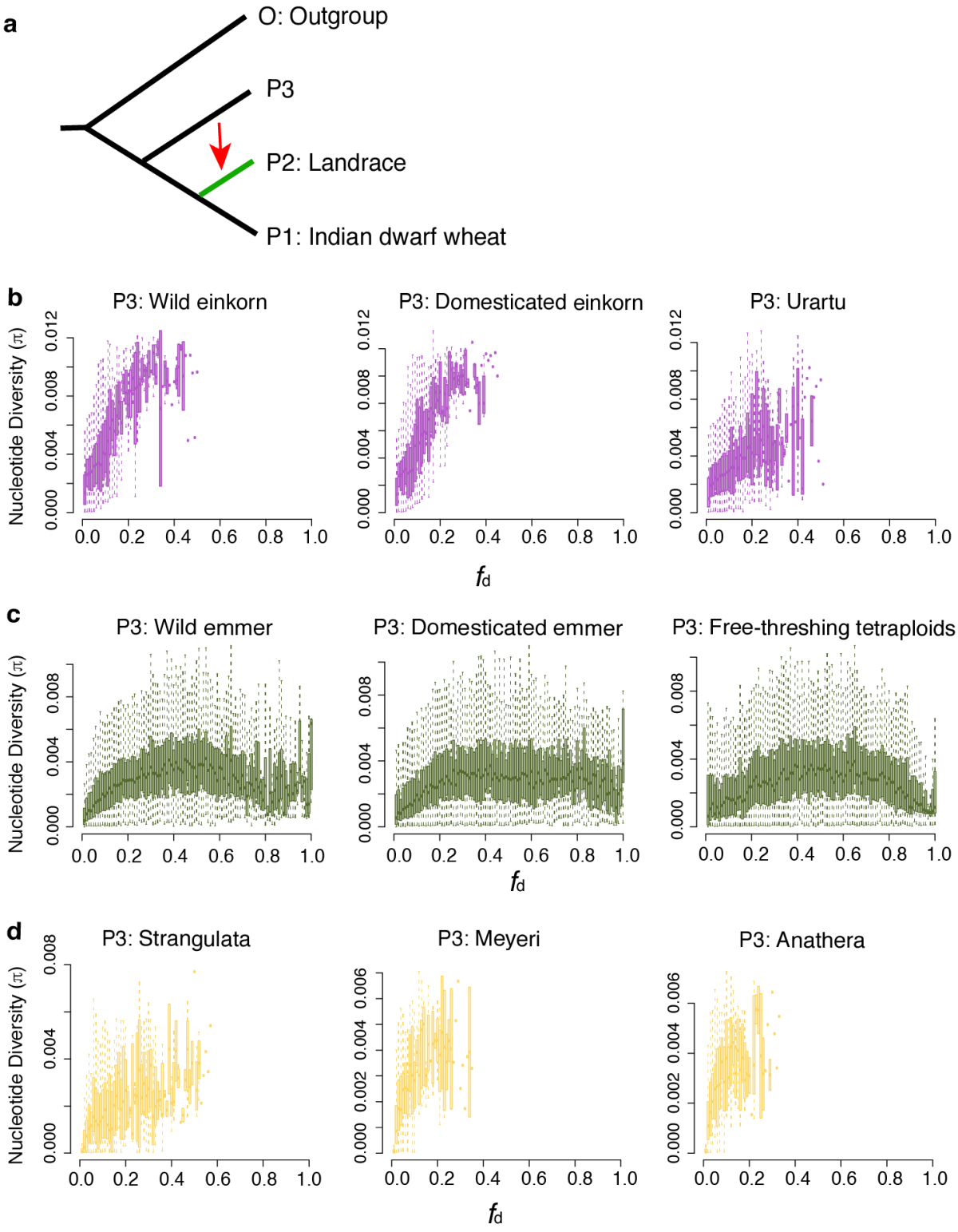


Supplementary Fig. 7. The distribution of window size of 100-SNP windows across the whole genome.

Supplementary Figure 8

Supplementary Fig. 8. Haplotype of specific introgression region on chromosome 2A. The haplotype of introgression is shown in the red rectangle. The haplotype on chromosome 2A from position 165Mb to 540Mb was shown for each accession. Each row represents a genomic position for all accessions, and the column represents the haplotype of each accession. The ploidy of each accession is presented at the bottom of the cluster tree with different colors. Dark pink represents homozygous genotype of alternative alleles, light pink represents heterozygous, and light yellow represents homozygous genotype of reference alleles.

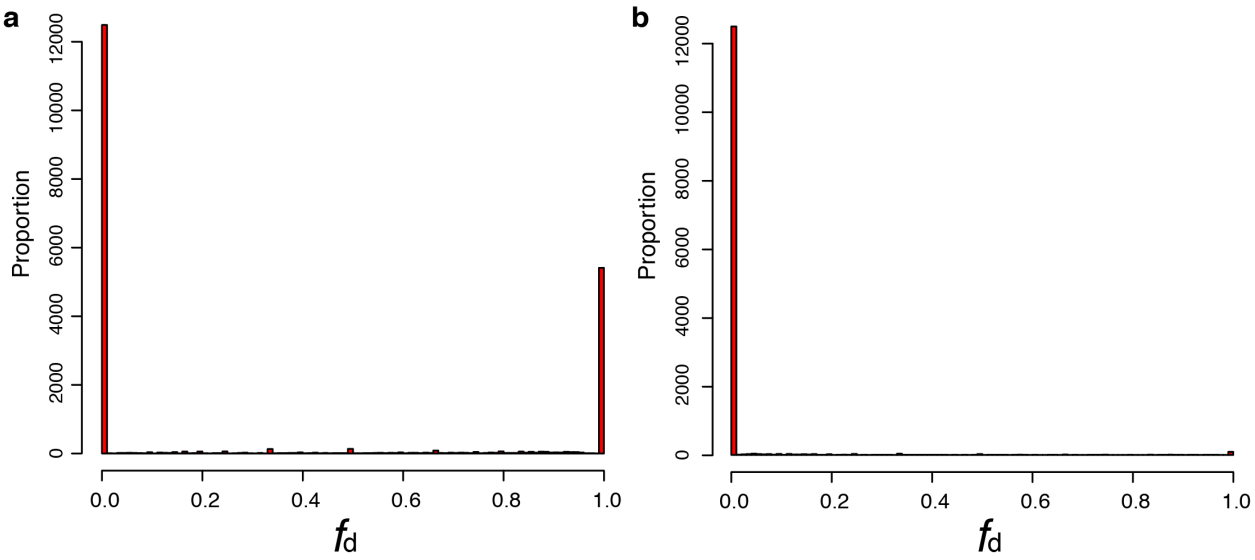
Supplementary Figure 9



Supplementary Fig. 9. Nucleotide diversity is a function of f_d. **a,** The topology used for

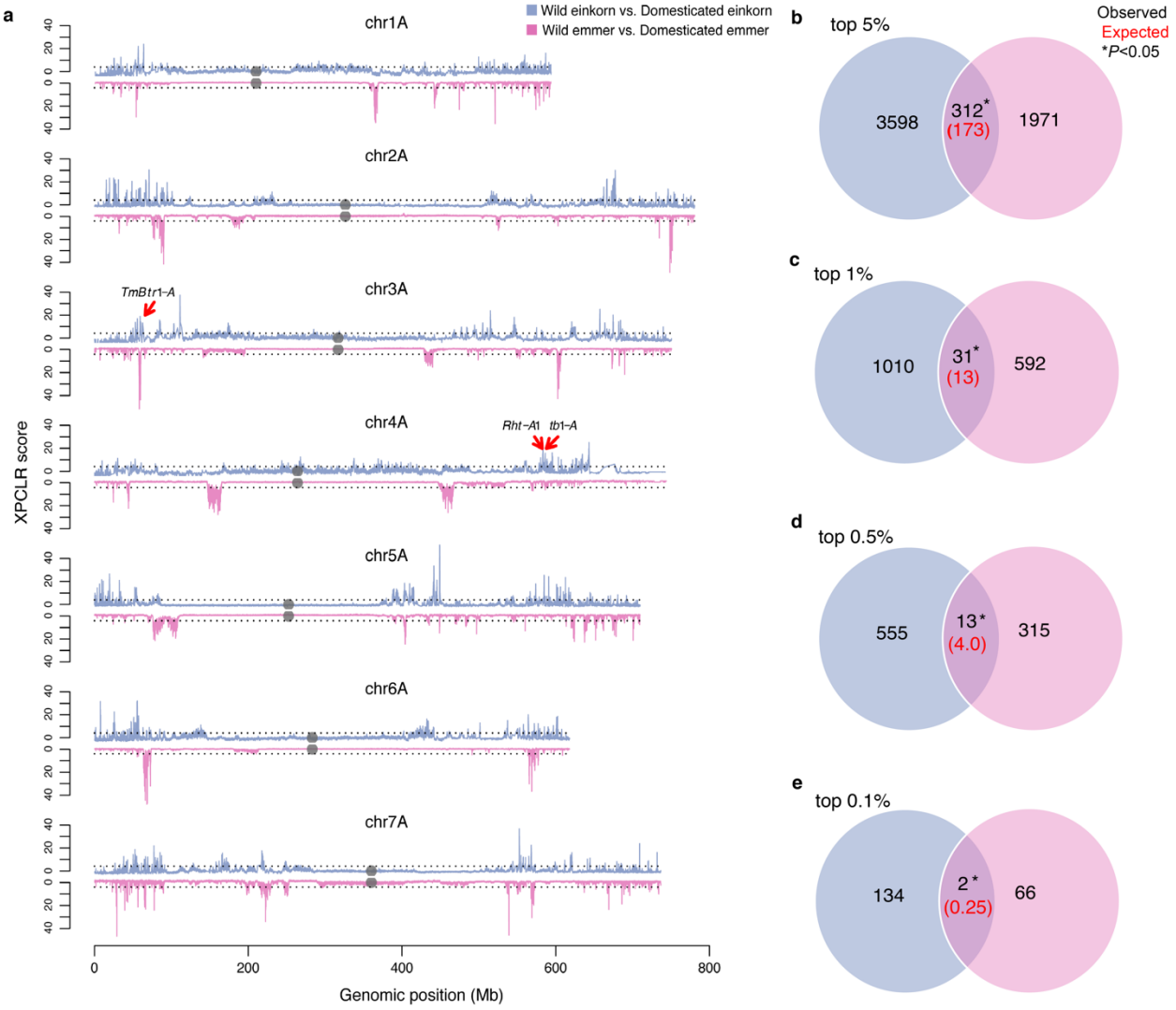
calculating f_d. **b,** Diploid species/subspecies in A lineage were used as P3. **c,** Tetraploid wheats were used as P3. **d,** Diploid species/subspecies in D lineage were used as P3.

Supplementary Figure 10



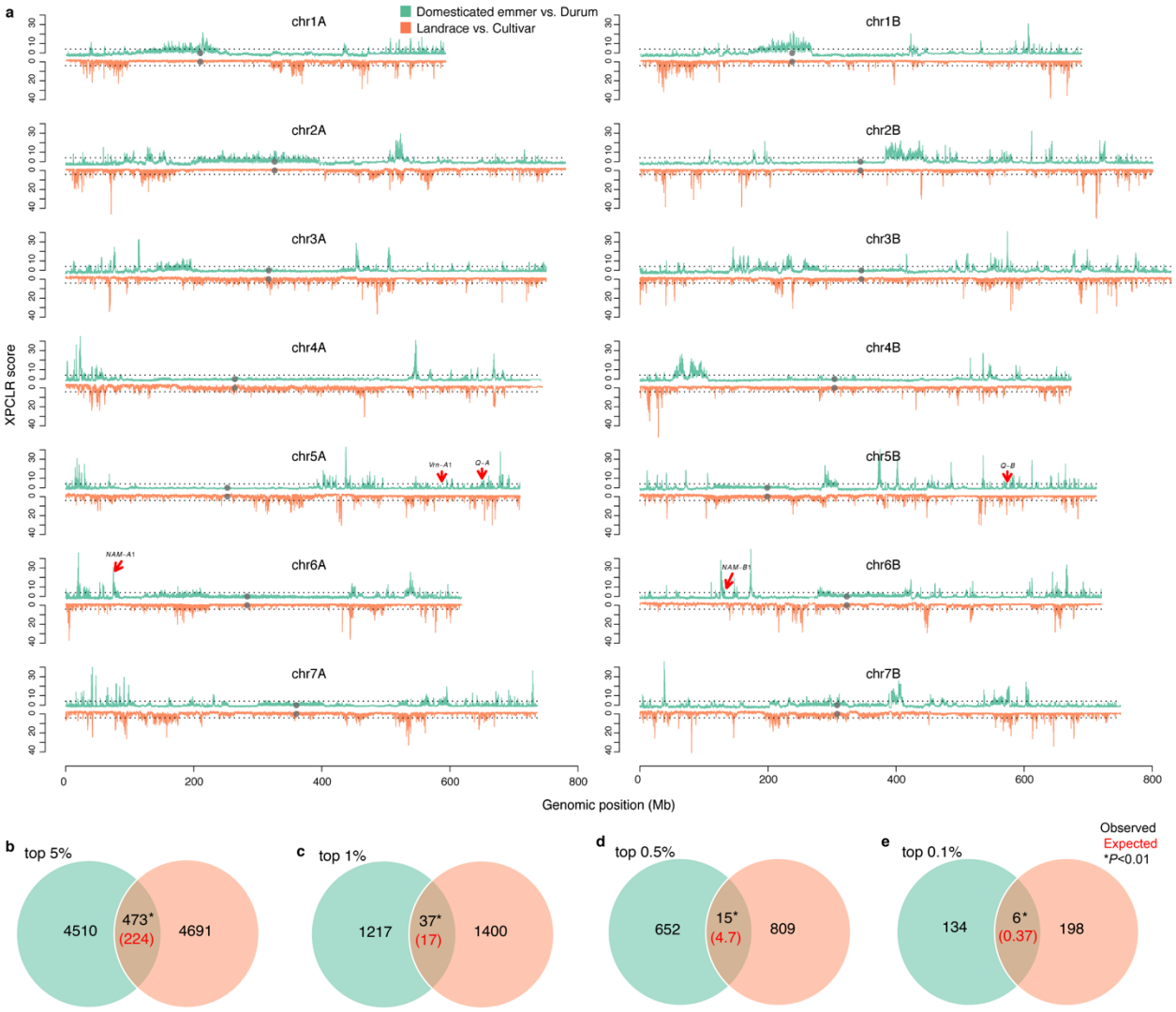
Supplementary Fig. 10. The distribution of f_d value of individual landraces. The f_d value for each landrace individual was obtained from the best donor (f_d = 1 and minimum IBS distance) of AB subgenome (a) or D subgenome (b).

Supplementary Figure 11



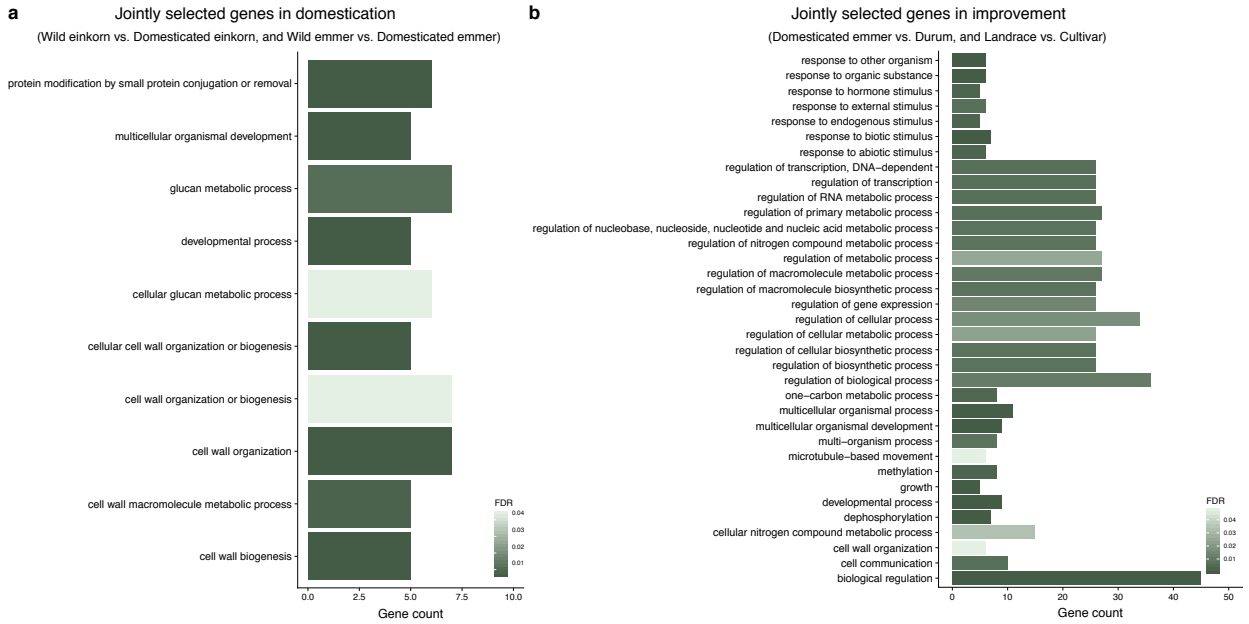
Supplementary Fig. 11. Comparison of selective sweeps detected from the paired domestication events. **a,** Selective sweeps across the genome. **b-e,** Number of jointly selected genes with different threshold of XP-CLR score from top 5%, top 1%, top 0.5%, and top 0.1%. Blue lines or circles represent the XP-CLR score of domestication from wild einkorn to domesticated einkorn. Pink lines or circles represent the XP-CLR score of domestication from wild emmer to domesticated emmer. Gray dots represent the positions of centromeres. Number in red brackets represent the theoretical number under the permutation test.

Supplementary Figure 12



Supplementary Fig. 12. Comparison of selective sweeps detected from the paired improvement events. **a,** Selective sweep across the genome. **b-e,** Number of jointly selected genes with different thresholds of top 5%, top 1%, top 0.5%, and top 0.1%. Green lines or circles represent the XP-CLR score of domestication from domesticated emmer to durum. Red lines or circles represent the XP-CLR score of domestication from landrace to cultivar. Gray dots represent the positions of centromeres. Number in red brackets represent the expected number under random conditions.

Supplementary Figure 13



Supplementary Fig. 13. GO enrichment for jointly selected genes. **a,** Enrichment for jointly selected genes in the paired domestication events. **b,** Enrichment for jointly selected genes in the paired improvement events. Only processes with five or more genes enriched were plotted.

Supplementary Figure 14

Supplementary Fig. 14. The non-BR phenotype of a synthetic hexaploid wheat (accession CWI76521). The CWI76521 was generated from hybridization between domesticated emmer accession PI272533 (AABB) with non-BR phenotype and *Ae. squarrosa* accession 518 (DD) with BR phenotype. Views from different sides are shown. Scale bars, 1 cm.

Supplementary references

1. (IWGSC), T. I. W. G. S. C. *et al.* Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, eaar7191 (2018).
2. Pont, C. *et al.* Tracing the ancestry of modern bread wheats. *Nat. Genet.* 1 (2019). doi:10.1038/s41588-019-0393-z
3. Ramu, P. *et al.* Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat. Genet.* **49**, 959–963 (2017).
4. Martin, S. H., Davey, J. W. & Jiggins, C. D. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* **32**, 244–257 (2015).
5. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
6. Faris, J. D. Wheat Domestication: Key to Agricultural Revolutions Past and Future. in *Genomics of Plant Genetic Resources: Volume 1. Managing, sequencing and mining genetic resources* (eds. Tuberosa, R., Graner, A. & Frison, E.) 439–464 (Springer Netherlands, 2014). doi:10.1007/978-94-007-7572-5_18
7. Salamini, F., Özkan, H., Brandolini, A., Schäfer-Pregl, R. & Martin, W. Genetics and geography of wild cereal domestication in the near east. *Nat. Rev. Genet.* **3**, 429–441 (2002).
8. Waisel, Y. Evolution of erect growth forms in domesticated wheats: possible effects of grazing. *Oecologia* **73**, 630–632 (1987).
9. Pourkheirandish, M. *et al.* Evolution of the Grain Dispersal System in Barley. *Cell* **162**, 527–539 (2015).
10. Lukaszewski, A. J. *et al.* A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science* **345**, 1251788 (2014).
11. Avni, R. *et al.* Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* **357**, 93–97 (2017).
12. Pourkheirandish, M. *et al.* On the origin of the non-brittle rachis trait of domesticated einkorn wheat. *Front. Plant Sci.* **8**, 2031 (2017).