Genomic insights into historical improvement of heterotic groups during modern hybrid maize breeding

Chunhui Li[1,12], Honghui Guan[1,12], Xin Jing[2,12], Yaoyao Li[3,12], Baobao Wang[4,12], Yongxiang Li[1], Xuyang Liu[1], Dengfeng Zhang[1], Cheng Liu[5], Xiaoqing Xie[5], Haiyan Zhao[6], Yanbo Wang[6], Jingbao Liu[7], Panpan Zhang[7], Guanghui Hu[8], Guoliang Li[8], Suiyan Li[9], Dequan Sun[9], Xiaoming Wang[1], Yunsu Shi[1], Yanchun Song[1], Chengzhi Jiao[2 ✉], Jeffrey Ross-Ibarra[10,11 ✉], Yu Li[1 ✉], Tianyu Wang[1 ✉] and Haiyang Wang[3 ✉]

Single-cross maize hybrids display superior heterosis and are produced from crossing two parental inbred lines belonging to genetically different heterotic groups. Here we assembled 1,604 historically utilized maize inbred lines belonging to various female heterotic groups (FHGs) and male heterotic groups (MHGs), and conducted phenotyping and genomic sequencing analyses. We found that the FHGs and MHGs have undergone both convergent and divergent changes for different sets of agronomic traits. Using genome-wide selection scans and association analyses, we identified a large number of candidate genes that contributed to the improvement of agronomic traits of the FHGs and MHGs. Moreover, we observed increased genetic differentiation between the FHGs and MHGs across the breeding eras, and we found a positive correlation between increasing heterozygosity levels in the differentiated genes and heterosis in hybrids. Furthermore, we validated the function of two selected genes and a differentiated gene. This study provides insights into the genomic basis of modern hybrid maize breeding.

M aize (*Zea mays* ssp. *mays* L.) is an important crop cultivated across the world, accounting for approximately 35.7% of the total worldwide cereal production[1]. Since the 1960s, farmers worldwide have primarily grown single-cross hybrid varieties that are the F_1 of a cross between two inbred lines belonging to two different heterotic groups. These single-cross hybrids display superior heterosis or hybrid vigour when compared with their inbred parents but more uniformity than open-pollinated varieties. Remarkably, maize yields in the central Corn Belt of the United States have increased over sevenfold during the past century (from 1,287 kg ha⁻¹ in 1930 to 9,595 kg ha⁻¹ in 2010), owing to the combined effects of continuous breeding of advanced hybrids, utilization of synthetic fertilizers and pesticides, and improvement of management practices (for example, high-density planting and mechanical harvesting)[2–4]. Although lagging behind the United States, similar trends are reported in other countries[5].

Two factors that have contributed critically to the success of single-cross hybrid maize breeding are the classification of elite germplasm into heterotic groups and the identification of high-yielding heterotic patterns. Initially, US corn breeders empirically grouped the parental lines with higher kernel yield and smaller tassels (such as B14, B37 and B73) into the female heterotic group (FHG) termed Stiff Stalk (SS), and the corresponding parents with

more pollen and longer pollen shedding duration (such as Mo17 and Oh43) were classified into the male heterotic group (MHG) termed Non-Stiff Stalk (NSS)[6]. Over the past few decades, these groups have expanded dramatically by the continuous addition of new inbred lines, accompanied by the formation of new heterotic groups, such as Iodent in the United States and Sipingtou (SPT), PA and PB in China[7,8]. Generally, new inbred lines were generated through making crosses within the same heterotic group with occasional introgressions from exotic germplasm, and elite inbred lines were selected on the basis of hybrid performance through crossing with tester lines from the opposite heterotic group, following the general guidance of "heterotic patterns" with pronounced heterosis[9,10]. Nowadays, the germplasm pools widely used in temperate maize breeding, such as SS and PA, are generally classified into the FHGs, whereas NSS, SPT and PB are generally classified into the MHGs. Meanwhile, SS×NSS and PA×SPT have been adopted as the predominant heterotic patterns widely used in the United States and China, respectively.

Previously, a few studies have investigated historical morphological changes of inbred lines during modern maize breeding and found convergent changes in several agronomic traits (such as reduced relative ear height (EP), fewer tassel branches, more vertical leaf angle and reduced anthesis-to-silking interval (ASI)) in the

[1]Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China. [2]Novogene Bioinformatics Institute, Beijing, China. [3]Guangdong Laboratory for Lingnan Modern Agriculture, State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, South China Agricultural University, Guangzhou, China. [4]Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing, China. [5]Institute of Food Crops, Xinjiang Academy of Agricultural Sciences, Urumqi, China. [6]Institute of Maize Research, Liaoning Academy of Agricultural Sciences, Shenyang, China. [7]Institute of Cereal Crops, Henan Academy of Agricultural Sciences, Zhengzhou, China. [8]Institute of Maize Research, Heilongjiang Academy of Agricultural Sciences, Harbin, China. [9]Institute of Forage and Grassland Sciences, Heilongjiang Academy of Agricultural Sciences, Harbin, China. [10]Department of Evolution and Ecology, University of California, Davis, CA, USA. [11]Center for Population Biology and Genome Center, University of California, Davis, CA, USA. [12]These authors contributed equally: Chunhui Li, Honghui Guan, Xin Jing, Yaoyao Li, Baobao Wang. ✉e-mail: jiaocz1990@163.com; rossibarra@ucdavis.edu; liyu03@caas.cn; wangtianyu@caas.cn; whyang@scau.edu.cn

parental inbred lines that enabled better adaptation of the inbred lines and their hybrids to high-density planting[11–14]. In addition, it has been documented that the female parental lines are often selected to have higher grain yield and smaller tassels, to increase hybrid seed production and reduce seed production cost[6,11]. The recent advances in molecular biology and genomics have instigated numerous efforts to explore the genomic bases of inbred line improvement and heterotic group evolution during modern maize breeding[13–20]. These studies have identified numerous candidate genes responsible for the morphological changes of the inbred lines and have documented divergent evolution of the various heterotic groups. However, we still lack a systematic assessment of the effect and contribution of artificial selection on phenotypic improvement and the underlying genomic changes of FHGs and MHGs on a population scale for different heterotic patterns during modern hybrid maize breeding. In general, breeding of the female and male parents still largely remains an empirical, time-consuming and cost-inefficient process, and the outcomes are often unpredictable for yield performance, which hinders the efficiency of hybrid maize breeding.

Here we resequenced the genomes of 1,604 diverse inbred lines from different heterotic groups used in modern hybrid maize breeding across the world and evaluated their phenotypic performance across ten environments. We find that the FHGs and MHGs experienced both convergent and divergent changes for different sets of agronomic traits. We also document the signatures of genetic improvement and differentiation between the FHGs and MHGs. Our results provide genomic insights into the historical improvement of maize parental lines and lay a foundation for exploiting the molecular basis of maize heterosis.

Results

Genomic variation and population structure. To assess the genetic variation representative of modern hybrid maize breeding, we resequenced the genomes of 1,604 inbred lines spanning different historical periods (Extended Data Fig. 1a and Supplementary Table 1), including 1,227 elite inbreds from China, 344 inbreds from the United States and 33 inbreds from other countries. The Chinese inbreds are classified into two breeding eras on the basis of their date of release and the time of the Seed Law coming into force: 408 from Era I (before 2000) and 760 from Era II (after 2000). The US inbreds consist of 85 early public lines (Era I) and 239 commercial lines with expired Plant Variety Protection (Era II). Each line was sequenced to an average genome coverage of 7.5× (between 5.2× and 33.5×) and aligned to the B73 RefGen_v4, enabling the identification of more than 18 million high-quality single nucleotide polymorphisms (SNPs) (Supplementary Tables 2 and 3 and Methods).

We investigated the population structure present among our 1,604 lines using neighbour-joining trees, principal component analysis (PCA) and the Bayesian approach implemented in fastStructure (Fig. 1, Extended Data Fig. 1b,c and Supplementary Table 1). These lines could be divided into 11 subgroups that can be further classified into six groups representing major heterotic groups widely used in maize breeding programmes over the past few decades: SPT, PA, PB, SS, NSS and Iodent, consisting of 131, 217, 88, 235, 431 and 124 inbreds, respectively, plus an admixed group of 378 inbreds. SPT contains two subgroups: the SPT1 subgroup originated from the Chinese landraces of SPT and Huangsipingtou, and the SPT2 subgroup is derived from SPT1 (ref. [21]). Similarly, PA contains the PA1 and PA2 subgroups, the latter being derived from the former. Iowa Stiff Stalk Synthetic and Amargo had a close genetic relationship[7,22] and were combined into the SS group. Moreover, Lancaster1, Lancaster2 and Zi330 were genetically close to each other[23] and were integrated into the NSS group. Among the six major heterotic groups, PA and SS were predominantly used as FHGs, while PB, SPT, NSS and Iodent were predominantly used as MHGs (Fig. 1c).

Morphological changes in the FHGs and MHGs. To systematically evaluate historical changes of agronomic traits both within and between the FHGs and MHGs during modern hybrid maize breeding, we divided each heterotic group into two breeding eras: Era I (before 2000) and Era II (after 2000). Only six heterotic groups, including two female groups (PA and US_SS) and four male groups (PB, SPT, CN_NSS and US_NSS), had more than 20 inbreds for each breeding era and thus could be used for subsequent statistical analysis (Fig. 1c and Supplementary Table 4).

We compared the phenotypic values of 21 agronomic traits of the six heterotic groups across Era I and Era II (Fig. 2a–c, Extended Data Fig. 2 and Supplementary Table 5). We found that on average, the FHGs had significantly higher grain yield per plant (GYPP) and yield-related traits, including kernel weight per ear (KWPE), ear weight (EW), kernel number per row (KNPR), hundred kernel weight (HKW), hundred kernel volume (HKV), kernel length (KL), kernel width (KW), ear diameter (ED), ear length (EL), kernel row number (KRN), cob weight (CW) and kernel ratio (KR), than the MHGs in Era I (Fig. 2a–c and Extended Data Fig. 2d), while the MHGs on average had larger EP and tassel branch number (TBN) than the FHGs in Era I (Fig. 2a). Of these traits, similar differences have been reported previously for GYPP, EW, HKW, EP and TBN between the FHGs and MHGs[6,11]. From Era I to Era II, 10 of the 21 traits displayed significant changes (Wilcoxon test, $P < 0.05$) in the FHGs, while 18 of them displayed significant changes in the MHGs (Fig. 2a–c and Extended Data Fig. 2d), implying that the male parental lines might have undergone more extensive artificial selection during modern hybrid maize breeding. Among these traits, 15 exhibited change in the same direction in both the FHGs and MHGs from Era I to Era II, and thus they were defined as convergently selected traits. Of these, seven (Group 1) reached a significant level. These traits are reduction in days to anthesis (DA), days to silking (DS), ASI, ear height (EH), EP and TBN, and increase in KR (Fig. 2a and Extended Data Fig. 2a). The other eight traits (Group 2) showed increases in both the FHGs and the MHGs from Era I to Era II, but reached a significant level only in the MHGs (Fig. 2b and Extended Data Fig. 2b). These eight traits are increases in GYPP, KWPE, EW, KNPR, HKW, HKV, KL and KW. In addition, three traits (Group 3)—namely, ED, CW and KRN—showed significant changes in opposite directions in the FHGs (decreases) and MHGs (increases) from Era I to Era II; thus, they were defined as divergently selected traits (Fig. 2c and Extended Data Fig. 2c). Finally, three traits (Group 4), including plant height (PH), tassel length (TL) and EL, showed no significant changes in either the FHGs or the MHGs from Era I to Era II (Extended Data Fig. 2d). These observations suggest that the FHGs and MHGs have undergone both convergent and divergent phenotypic changes during modern hybrid maize breeding.

Accumulation of favourable alleles linked to morphological changes. To dissect the genetic basis of the respective trait improvements of the FHGs and MHGs, we performed genome-wide association studies (GWAS) for 21 important agronomic traits across all 1,604 inbred lines at five agro-ecologically distinct locations in China in 2018 and 2019. On the basis of a cut-off of $P < 10^{-6}$ for GWAS, we identified 2,360 significant associations at 1,847 SNPs (Supplementary Table 6). A total of 429 associated genes were identified for the 21 traits, of which 74 were detected for at least two traits (Supplementary Table 6), implying a pleiotropic role of these candidate genes. Among the 317 candidate genes identified, a total of 207, 56 and 24 candidate genes were significantly associated with the seven traits of Group 1, the eight traits of Group 2 and the three traits of Group 3, respectively. Among them, *ZmRap2.7* (Zm00001d010987), *DRL2* (Zm00001d048083), *Sxd1* (Zm00001d015985) and *BIF4* (Zm00001d037691) have been shown to regulate important agronomic traits in maize[24–27].
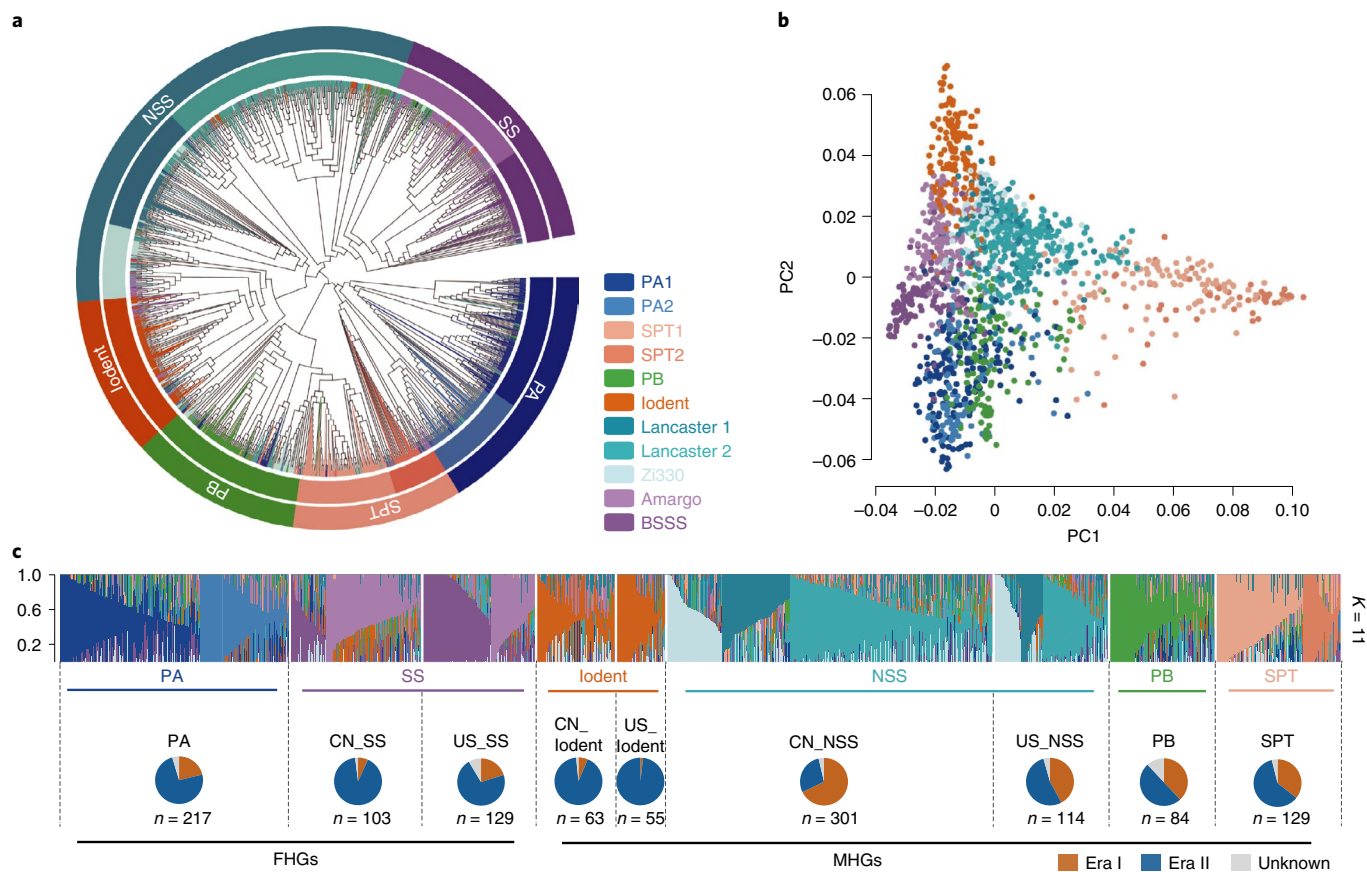
Fig. 1 | Population structure of 1,604 inbred lines. a, Phylogenetic tree of all inbreds. Correspondences between branches and main subgroups from fastStructure are indicated by colours. Eleven subgroups (inner circle) were merged into six major groups (outer circle). BSSS, Iowa Stiff Stalk Synthetic. **b**, PCA plot of the first two components for all inbreds. The dot colours correspond to the 11 subgroups of the phylogenetic tree. **c**, Diagram of inbred classification based on different hierarchies. From top to bottom: population structure when $K = 11$, six heterotic groups when considering the pedigree and breeding experience of 11 subgroups, nine heterotic groups when considering the origin/breeding eras of the inbreds, and two pools when considering breeding use in hybrid development (that is, the FHGs and MHGs).

To assess the relationships of these GWAS associations with the observed morphological changes in the FHGs and MHGs, we chose the associated SNPs ($P < 10^{-5}$) to investigate their allele frequency dynamics during modern breeding (Methods). Consistent with trends of trait improvement in the FHGs and MHGs, the frequency of favourable alleles at the associated SNPs showed convergent increases in both the FHGs and MHGs for the convergently improved traits (such as KR and GYPP) (Fig. 2d and Extended Data Fig. 3a). Similarly, the frequency of favourable alleles at the associated SNPs exhibited anti-directional changes in the FHGs (decreases) and the MHGs (increases) for the divergently changed traits (such as ED) (Fig. 2d and Extended Data Fig. 3a). We also found that the number of accumulated favourable alleles in the FHGs and MHGs correlated well with the phenotypic changes from Era I to Era II (Fig. 2e,f and Extended Data Fig. 3b,c). Together, these observations suggest that the accumulation of favourable alleles underlies agronomic trait improvement of the FHGs and MHGs during modern maize breeding.

Selection signals during modern hybrid maize breeding. Next, we used a cross-population likelihood method (the cross-population composite likelihood ratio (XP-CLR)) to identify the genome-wide regions targeted by selection during improvement of the FHGs and MHGs from Era I to Era II (Methods). We identified a total of 18,665 regions affected by artificial selection in at least one heterotic group (Fig. 3 and Supplementary Table 7). These genomic regions (average size 41.0 kilobases (kb)) together span about 767.4 megabases (Mb) (Supplementary Table 8), indicating that selection during maize breeding affected a substantial portion (~36.4%) of the maize genome. Notably, the improvement signals within each heterotic group were mainly distributed in intergenic regions, particularly in promoters and regions 10–50 kb distant from the gene bodies (Extended Data Fig. 4a,b), suggesting that selection may have preferentially targeted non-genic regulatory elements during heterotic group improvement. This finding is consistent with several previous studies highlighting the pivotal roles of structural variation in gene regulatory regions in regulating gene function and important agronomic traits in maize[28–31].

We identified a total of 16,993 genes in the selected regions, ranging from 2,863 to 4,355 genes within each individual heterotic group (Supplementary Table 8). Cross-heterotic group comparisons showed that among the 16,993 genes, 2,820 (16.6%) were shared between at least one FHG and at least one MHG (Fig. 3 and Supplementary Table 9), 1,630 (9.6%) were specifically selected in at least two MHGs, and 354 (2.1%) were specifically selected in the two FHGs (Supplementary Table 9). Gene Ontology (GO) enrichment analysis revealed that 111 GO terms were shared among at least two heterotic groups (Supplementary Table 10). Among them, a total of 54 GO terms were shared between at least one FHG and one MHG, 51 were shared in at least two MHGs, and 6 were shared in two FHGs (Supplementary Table 10). Statistical analysis showed that the sharing degree of selected genes and GO terms between
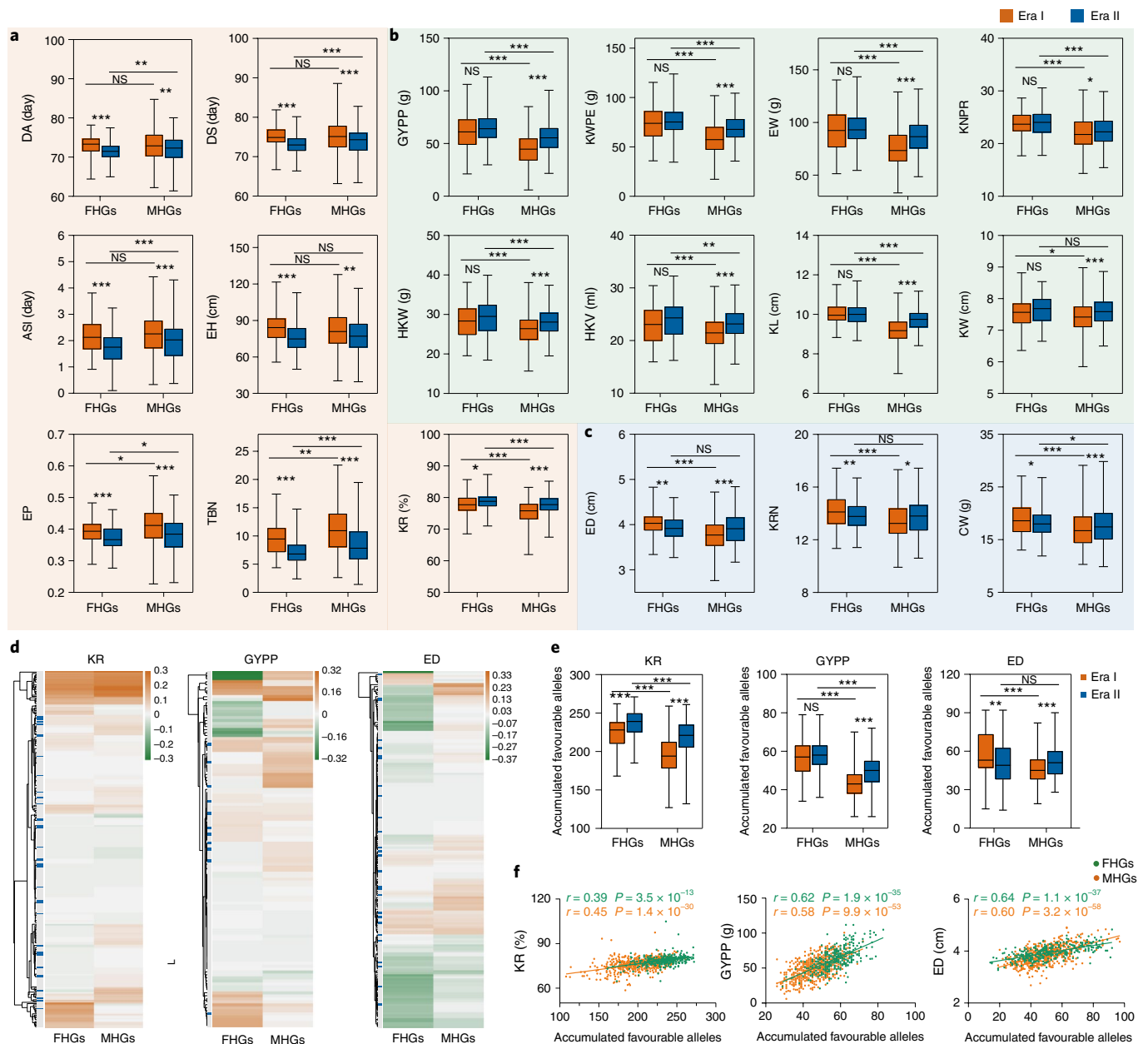
Fig. 2 | Morphological trait improvement and accumulation of favourable alleles in the FHGs and MHGs during modern hybrid maize breeding.
a–c, Phenotypic changes in the FHGs and MHGs across different breeding eras. The traits of Group 1 (**a**) include DA, DS, ASI, EH, EP, TBN and KR. The traits of Group 2 (**b**) include GYPP, KWPE, EW, KNPR, HKW, HKV, KL and KW. The traits of Group 3 (**c**) include ED, KRN and CW. The P values of FHGs[Era I] versus FHGs[Era II], MHGs[Era I] versus MHGs[Era II], FHGs[Era I] versus MHGs[Era I] and FHGs[Era II] versus MHGs[Era II] by two-sided Wilcoxon tests are presented in Supplementary Table 5. **d**, Profile of favourable allele frequency changes for KR, GYPP and ED at the GWAS associated SNPs in the FHGs and MHGs. Orange indicates an increase, while green indicates a decrease in the frequency of a favourable allele from Era I to Era II. Each row represents an associated SNP. Blue and grey (in the first column) mark rows representing associated SNPs at the thresholds of $P < 10^{-6}$ and $P < 10^{-5}$, respectively. **e**, The number of favourable alleles accumulated for KR, GYPP and ED in the FHGs and MHGs across different breeding eras. The P values of FHGs[Era I] versus FHGs[Era II], MHGs[Era I] versus MHGs[Era II], FHGs[Era I] versus MHGs[Era I] and FHGs[Era II] versus MHGs[Era II] by two-sided Wilcoxon tests are as follows: for KR, $P = 6.46 \times 10^{-6}$, 5.89×10^{-29}, 7.27×10^{-17} and 2.83×10^{-22}, respectively; for GYPP, $P = 2.29 \times 10^{-1}$, 1.00×10^{-19}, 8.49×10^{-20} and 1.27×10^{-27}, respectively; and for ED, $P = 1.30 \times 10^{-3}$, 8.77×10^{-6}, 4.08×10^{-8} and 2.65×10^{-1}, respectively. **f**, Correlation between the number of accumulated favourable alleles and improved traits for KR, GYPP and ED in the FHGs and MHGs. The Pearson correlation coefficient (r) and P value are presented. For the box plots in **a–c** and **e**, the central lines show the medians, the box limits indicate the 25th and 75th percentiles, and the whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles. The exact sample sizes (n) are 72, 253, 330 and 267 for FHGs[Era I], FHGs[Era II], MHGs[Era I] and MHGs[Era II], respectively, in **a–c** and **e**. Significant differences are indicated: ***$P < 0.001$; **$P < 0.01$; *$P < 0.05$. NS, not significant.

the FHGs and MHGs or within the FHGs/MHGs was significantly higher than the background level (Extended Data Fig. 4c), suggesting that convergent selection in the FHGs and/or MHGs occurred at both the gene and pathway levels.

To search for candidate genes underlying the observed convergent and divergent phenotypic improvement between the FHGs and MHGs, we identified 589 genes (of the 2,820 shared genes) containing 1,562 non-synonymous SNPs exhibiting co-directional changes
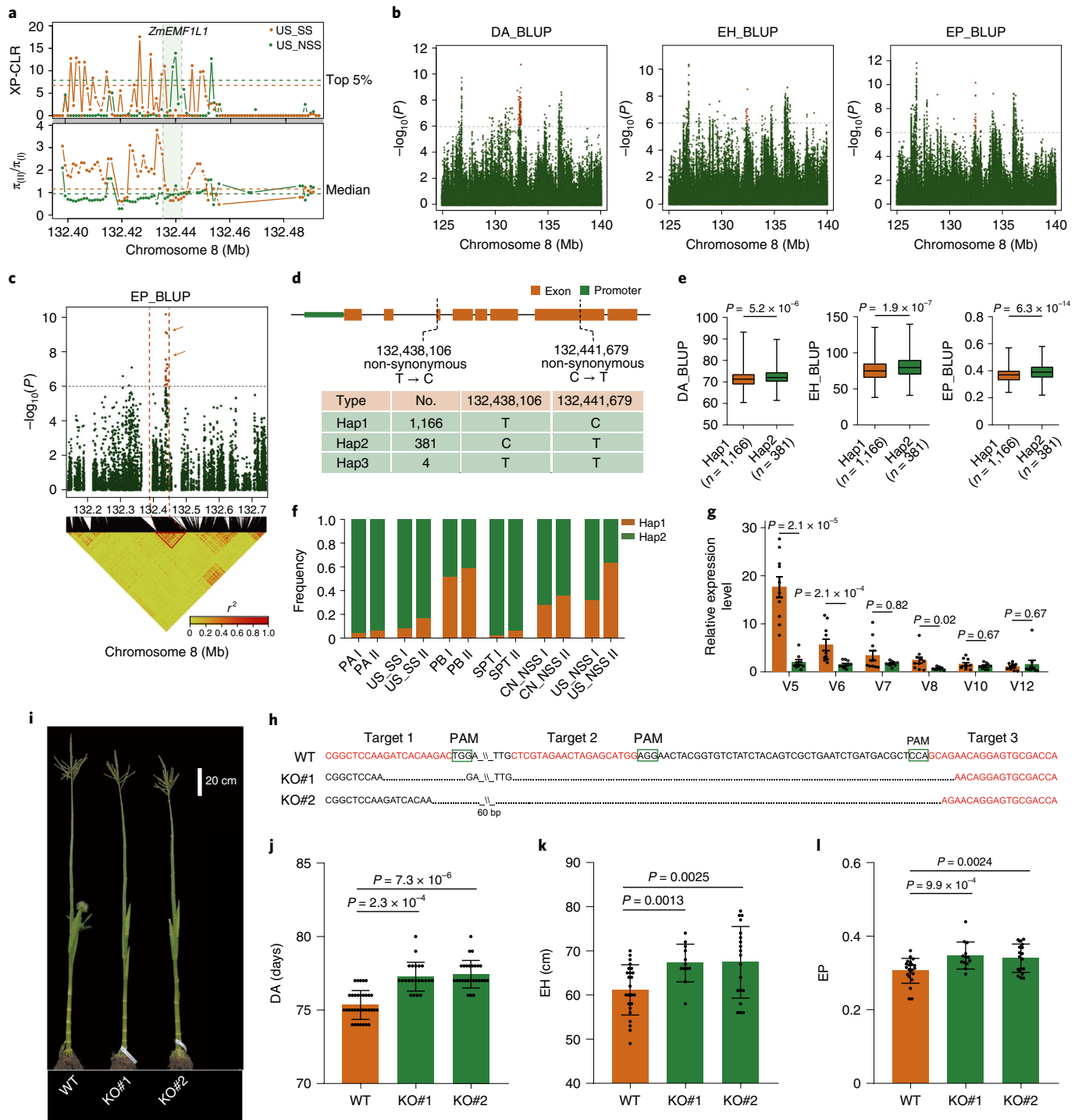
Fig. 3 | Atlas of selection signatures in the FHGs and MHGs. Genome-wide selective signals (XP-CLR scores) in each of the six heterotic groups. The serial numbers of the chromosomes (1–10) are marked along the *x* axes. The horizontal red dashed lines represent the cut-offs that define statistical significance. Representative selected genes (selected in at least two heterotic groups) with known or putative functions in regulating plant growth and development and abiotic stress responses are marked above the selective signal peaks. The genes in green were selected in both the FHGs and MHGs, and the genes in orange were selected in either the FHGs only or the MHGs only.

in allele frequencies between the FHGs and MHGs, and 28 genes containing 46 non-synonymous SNPs exhibiting anti-directional changes in allele frequencies between the FHGs and MHGs (Extended Data Fig. 4d,e and Supplementary Table 11). In addition, from the 1,630 genes selected only in the MHGs, we identified 400 genes containing 857 non-synonymous SNPs exhibiting similar increases or reductions in allele frequencies in the MHGs, but not in the FHGs (Extended Data Fig. 4f and Supplementary Table 11). Among them, dozens of genes were reported to regulate flowering time, plant architecture, kernel development and abiotic stress responses (Extended Data Fig. 5 and Supplementary Table 12). Notably, five undescribed genes (Zm00001d039919, Zm00001d027925, Zm00001d042314, Zm00001d010894 and Zm00001d025992) overlapped with the GWAS genes associated with the traits with observed convergent changes; these five could thus be viewed as high-confidence candidate genes underlying agronomic improvement in the parental inbred lines.

Functional validation of two selected genes. To assess the biological relevance of the identified candidate genes selected in the FHGs and MHGs, we selected two candidate genes for functional validation. Zm00001d010894 was identified as a commonly selected gene in US_SS (FHG) and US_NSS (MHG) (Fig. 4a) and was associated with three agronomic traits: DA, EH and EP (Fig. 4b). It encodes a protein homologous to the *Arabidopsis* EMBRYONIC FLOWER1 (EMF1) protein, which is involved in the control of flowering time and plant architecture[32], and thus was named *Zea mays EMF1-like* 1 (*ZmEMF1L1*). *ZmEMF1L1* was mapped in a genomic region from 132.36 to 132.45 Mb, which contained 14 significantly associated

SNPs (Fig. 4c). Two non-synonymous SNPs significantly associated with all three traits formed two major haplotypes (Fig. 4d). The accessions with Hap1 showed significantly shorter DA and lower EH and EP than those with Hap2. Thus, Hap1 was deemed the favourable haplotype (Fig. 4e). The frequency of Hap1 simultaneously increased in all heterotic groups (both FHGs and MHGs) during modern maize breeding (Fig. 4f). Gene expression analysis revealed that Hap1 had a significantly higher expression level than Hap2 (Fig. 4g), suggesting that higher expression of *ZmEMF1L1* confers shorter DA and lower EH and EP. Consistent with this deduction, two independent knockout lines (KO#1 and KO#2) of *ZmEMF1L1* generated via the CRISPR–Cas9 technology displayed delayed flowering and increased EH and EP compared with the wild type (Fig. 4h–l), thus validating *ZmEMF1L1* as a likely selective target for improving flowering time and plant architecture in various heterotic groups of maize.

Zm00001d025992 was detected as a candidate gene selected only in two MHGs: SPT and US_NSS (Fig. 5a). It was significantly associated with multiple kernel-related traits (KL, HKW and HKV) (Fig. 5b) and encodes a protein of unknown function. It was designated *ZmKW10* (located on chromosome 10) and contained 41 significantly associated SNPs (Fig. 5c). Two non-synonymous SNP variants significantly associated with HKV formed two major haplotypes (Fig. 5d). The accessions carrying Hap1 showed significantly higher KL, HKW and HKV than those with Hap2, and thus Hap1 was deemed the favourable haplotype (Fig. 5e). Notably, its frequency increased in all four MHGs, but not in either of the two FHGs (Fig. 5f). Expression analysis revealed that Hap1 displayed significantly lower expression than Hap2 during kernel development (Fig. 5g),

Fig. 4 | Identification and functional validation of *ZmEMF1L1*. a, XP-CLR (top) and ratio of nucleotide diversity (π) (bottom) between Era II and Era I. The candidate gene *ZmEMF1L1* is shown with green shading. **b**, Manhattan plots for DA, EH and EP on chromosome 8. The orange dots represent the significantly associated signals around the candidate gene *ZmEMF1L1* at the threshold of $P < 1.0 \times 10^{-6}$. BLUP, best linear unbiased predictor. **c**, Local Manhattan plot (top) and linkage disequilibrium (LD) heat map (bottom). The candidate region lies between the orange dashed lines. The arrows indicate the two significantly associated non-synonymous SNPs (132,438,106 and 132,441,679) located in the genic region of *ZmEMF1L1*. **d**, Gene structure and haplotype analyses of *ZmEMF1L1*. **e**, Box plots for DA, EH and EP for the two haplotypes. In each plot, the centre line indicates the median, the box limits indicate the upper and lower quartiles, and the whiskers indicate 1.5× the interquartile range; *n* indicates the number of inbred lines with the same haplotype. The significance of the difference was analysed using a two-sided Wilcoxon test. **f**, Haplotype frequency changes for *ZmEMF1L1* in different breeding eras of the six heterotic groups. **g**, Comparison of *ZmEMF1L1* expression levels between inbred lines carrying Hap1 (orange) and Hap2 (green) in leaf tissue of six different developmental stages, detected by quantitative real-time PCR. The data are presented as mean ± s.d. (*n* = 10 inbred lines with three technical replicates). Statistical significance was determined using two-sided Wilcoxon tests. **h**, Knockout of *ZmEMF1L1* by CRISPR–Cas9 technology. PAM, protospacer adjacent motif. **i–l**, Images (**i**) and statistics for DA (**j**), EH (**k**) and EP (**l**) of wild type (WT) and CRISPR-knockout (KO#1 and KO#2) plants. The dots indicate individual plants ($n_{WT} = 34$, $n_{KO\#1} = 22$, $n_{KO\#2} = 27$ in **j**; $n_{WT} = 24$, $n_{KO\#1} = 12$, $n_{KO\#2} = 19$ in **k**; and $n_{WT} = 21$, $n_{KO\#1} = 13$, $n_{KO\#2} = 19$ in **l**). The data are presented as mean ± s.d. The *P* values of one-sided *t*-tests are shown.

Fig. 5 | Identification and functional validation of *ZmKW10*. a, XP-CLR (top) and ratio of π (bottom) of *ZmKW10*. The candidate gene is shown with green shading. **b**, Manhattan plots for KL, HKW and HKV on chromosome 10. The orange dots represent the significantly associated signals around the candidate gene *ZmKW10* at the threshold of $P < 1.0 \times 10^{-6}$. HRB, Harbin. **c**, Local Manhattan plot (top) and LD heat map (bottom). The candidate region lies between the orange dashed lines. The arrows indicate the two significantly associated non-synonymous SNPs (135,578,575 and 135,578,782) located in the genic region of *ZmKW10*. **d**, Gene structure and haplotype of *ZmKW10*. **e**, Box plots for KL, HKW and HKV for the two haplotypes. In each plot, the centre line indicates the median, the box limits indicate the upper and lower quartiles, and the whiskers indicate 1.5× the interquartile range; *n* indicates the number of accessions with the same haplotype. The significance of the difference was analysed using a two-sided Wilcoxon test. **f**, Haplotype frequency for *ZmKW10* in different breeding eras of six heterotic groups. **g**, Comparison of *ZmKW10* expression levels between inbred lines carrying Hap1 (orange) and Hap2 (green) at three key stages of kernel development (6, 12 and 18 days after pollination (DAP)), detected by quantitative real-time PCR. The data are presented as mean ± s.d. (*n* = 10 inbred lines with three technical replicates). Statistical significance was determined using two-sided Wilcoxon tests. **h**, Knockout of *ZmKW10* by CRISPR–Cas9 technology. **i–l**, Images (**i**) and statistics for KL (**j**), KW (**k**) and HKW (**l**) of WT and CRISPR-knockout (KO#1 and KO#2) plants. The dots indicate individual ears ($n_{WT} = 24$, $n_{KO\#1} = 21$, $n_{KO\#2} = 15$) in **j–l**. The data are presented as mean ± s.e.m. The *P* values of one-sided *t*-tests are shown.

suggesting a repressive role of *ZmKW10* in regulating kernel size and weight. In support of this notion, knocking out *ZmKW10* using CRISPR–Cas9 technology caused significantly increased KL and HKW (Fig. 5h–l). We therefore concluded that *ZmKW10* may be a preferred selective target for grain yield improvement in the MHGs.

Differentiation between the FHGs and MHGs. Differentiation between the FHGs and MHGs in heterotic patterns is of great importance for hybrid breeding[33–36]. We thus chose PA (FHG) × SPT (MHG) and US_SS (FHG) × US_NSS (MHG), two heterotic patterns dominant in China and the United States, respectively, to investigate the degree and pattern of genetic differentiation between the FHGs and MHGs. We measured the pairwise fixation statistic (F_{ST}) and allele frequency difference (AFD) in these heterotic groups. The average F_{ST} values for PA × SPT and US_SS × US_NSS were 0.20 and 0.13, respectively, similar to that of Dent × Flint (0.14, a heterotic pattern widely used in Europe)[19], but higher than that between maize and teosinte[37], indicating that there is strong population differentiation between the female and male groups. Notably, we observed that the average F_{ST} values for both the PA × SPT and US_SS × US_NSS patterns increased significantly from Era I to Era II (Fig. 6a), indicating that modern breeding efforts have continuously driven genomic differentiation between the female and male groups, as previously noted[16]. On the basis of the top 5% of both F_{ST} and AFD, 1,483 and 1,639 differentiated genomic regions were identified for PA × SPT and US_SS × US_NSS, respectively (Supplementary Table 13), which covered 78.5 Mb and 81.7 Mb of the assembled genome; individual differentiated regions had mean sizes of 52.9 kb and 49.8 kb, respectively (Supplementary Table 14). Within the above differentiated regions, we identified 1,807 and 2,215 genes for PA × SPT and US_SS × US_NSS, respectively (Supplementary Table 15). Strikingly, only 117 overlapping genes were identified between the two heterotic patterns. When compared with the 1,407 genes differentially selected in Dent × Flint[19], only 103 and 28 overlapping genes were identified for PA × SPT and US_SS × US_NSS, respectively. These results suggest that artificial selection and breeding improvement of inbreds for different heterotic patterns targeted largely different genomic regions/genes during modern hybrid maize breeding. In addition, when compared with the selected genes obtained by XP-CLR, we found that only 483 of the 1,807 differentiated genes for PA × SPT overlapped with the selected genes in the PA or SPT group, and only 836 of the 2,215 differentiated genes for US_SS × US_NSS overlapped with the selected genes in the US_SS or US_NSS group. Moreover, analysis of allele frequency changes (based on non-synonymous SNPs) from Era I to Era II revealed that for the PA × SPT pattern, 20 of the 483 overlapping genes were selected in both parental pools, and 226 were selected only in one of the parental pools. Similarly, 82 of the 836 overlapping genes were selected in both parental pools and 464 were selected only in one of parental pools during maize breeding in the US_SS × US_NSS pattern (Supplementary Table 16). Furthermore, when compared with the 317 GWAS associated genes ($P < 10^{-6}$ cut-off), only 19 and 26 overlapping genes were identified with the 1,807 differentiated genes for PA × SPT and the 2,215 differentiated genes for US_SS × US_NSS, respectively, suggesting that the majority of the differentiated genes perhaps contribute to variation in traits not sampled in our GWAS or undetected because of falling short of the threshold of our GWAS analysis.

To identify the genes that have been subjected to continuous differential selection between the FHGs and MHGs during modern breeding, we combined the differentiated regions detected in China (PA × SPT, PA I × SPT I and PA II × SPT II) and in the United States (US_SS × US_NSS, US_SS I × US_NSS I and US_SS II × US_NSS II). We found that 694 genes for PA × SPT and 542 genes for US_SS × US_NSS have been under continuous differential selection in both Era I and Era II (Fig. 6b), of which 478 and 375 genes contained

2,563 and 2,136 non-synonymous variants and showed significant allele frequency changes in opposite directions between the female and male groups of PA × SPT and US_SS × US_NSS, respectively (Supplementary Table 17). These include a number of genes with known or putative functions in regulating plant growth and development and abiotic stress responses (Supplementary Table 18), such as *DIL1* (Zm00001d038087)[38], *ZmHY2* (Zm00001d011876)[39], *ZMM15* (Zm00001d013259)[40], *SBE1* (Zm00001d014844)[41] and *LYCE1* (Zm00001d011210)[42] (Extended Data Fig. 6).

To explore the potential roles of the differentiated genes in regulating heterosis in the hybrids, we designed four F_1 test-cross populations, CNH3754 (a PA inbred) × 88 SPT inbreds (yielding 88 hybrids), 91 PA inbreds × Jing2416 (an SPT inbred) (yielding 91 hybrids), Xunshi104-8 (an SS inbred) × 106 US_NSS inbreds (yielding 106 hybrids) and 101 US_SS inbreds × F62 (an NSS inbred) (yielding 101 hybrids), and investigated the relationship between the heterozygosity levels of the hybrids (on the basis of non-synonymous SNPs in the differentiated genes) and their better parent heterosis (BPH) of GYPP, EW and KWPE. We found no significant correlation between heterozygosity levels and any of the traits on the basis of randomly selected sets of non-synonymous SNPs except KWPE and EW in the CNH3754 × SPT population (Fig. 6c–f and Extended Data Fig. 7a–d); however, we detected modest but significant positive correlations between heterozygosity at non-synonymous SNPs in the 478 and 375 differentiated genes and BPH for all traits in PA × SPT and US_SS × US_NSS, respectively (except the PA × Jing2416 population) (Fig. 6g–j and Extended Data Fig. 7a–d). In addition, we found significant positive correlations between the number of accumulated superior heterozygous alleles and heterosis levels in their hybrids (assessed by yield traits) (Fig. 6k–n and Extended Data Fig. 7a–d). These observations suggest that heterozygosity at these differentiated genes may play a role in promoting yield heterosis in the hybrids.

Functional validation of a candidate differentiated gene. We next selected one differentiated candidate gene (Zm00001d039284) of the PA × SPT heterotic pattern for functional validation (Extended Data Fig. 8a). Zm00001d039284 encodes a glycosyltransferase-like protein, which is highly homologous to *Arabidopsis ABA INSENSITIVE 8* (*ABI8*)/*ELONGATION EFFECTIVE 1* (*ELD1*)/*KOBITO1* (*KOB1*)[43]. We therefore named it *Zea mays KOBITO1* (*ZmKOB1*). Plant glycosyltransferases transfer the sugars from the activated sugar donors to acceptor molecules (such as plant hormones, peptides, proteins and secondary metabolites) and are known to play an important role in regulating plant development and stress responses[44,45]. Two non-synonymous SNP variants in the coding region of *ZmKOB1* formed four major haplotypes (Extended Data Fig. 8b). The accessions carrying Hap1 showed significantly higher KWPE, KNPR and EL than those with Hap2 (Extended Data Fig. 8c). Notably, the frequencies of Hap1 and Hap2 displayed opposite trends of changes in the PA and SPT heterotic groups (Extended Data Fig. 8d). To confirm the effect of this gene on grain-yield-related traits, we generated *Zmkob1* knockout mutants and overexpression lines in the maize inbred line ZC01 background. Two independent knockout lines (KO#1 and KO#2) and two overexpression lines (OE3 and OE4) were analysed in the T_3 generation (Fig. 7a–i). Compared with the non-transgenic control plants, the knockout lines had reduced KWPE, KNPR and EL (Fig. 7b–e), whereas the overexpression lines (OE3 and OE4) exhibited the opposing phenotypes (Fig. 7f–i). To further test whether *ZmKOB1* contributes to yield heterosis, we generated F_1 hybrids of *ZmKOB1–OE4* (or its null segregant line, CK4) with Zheng58 (Z58, an elite inbred line). We found that the Z58/OE4 hybrids had significantly higher KNPR, KWPE and BPH than the Z58/CK4 hybrids (Fig. 7j–n). Taken together, these findings suggest that *ZmKOB1* probably plays a role in promoting yield heterosis through regulating ear development.
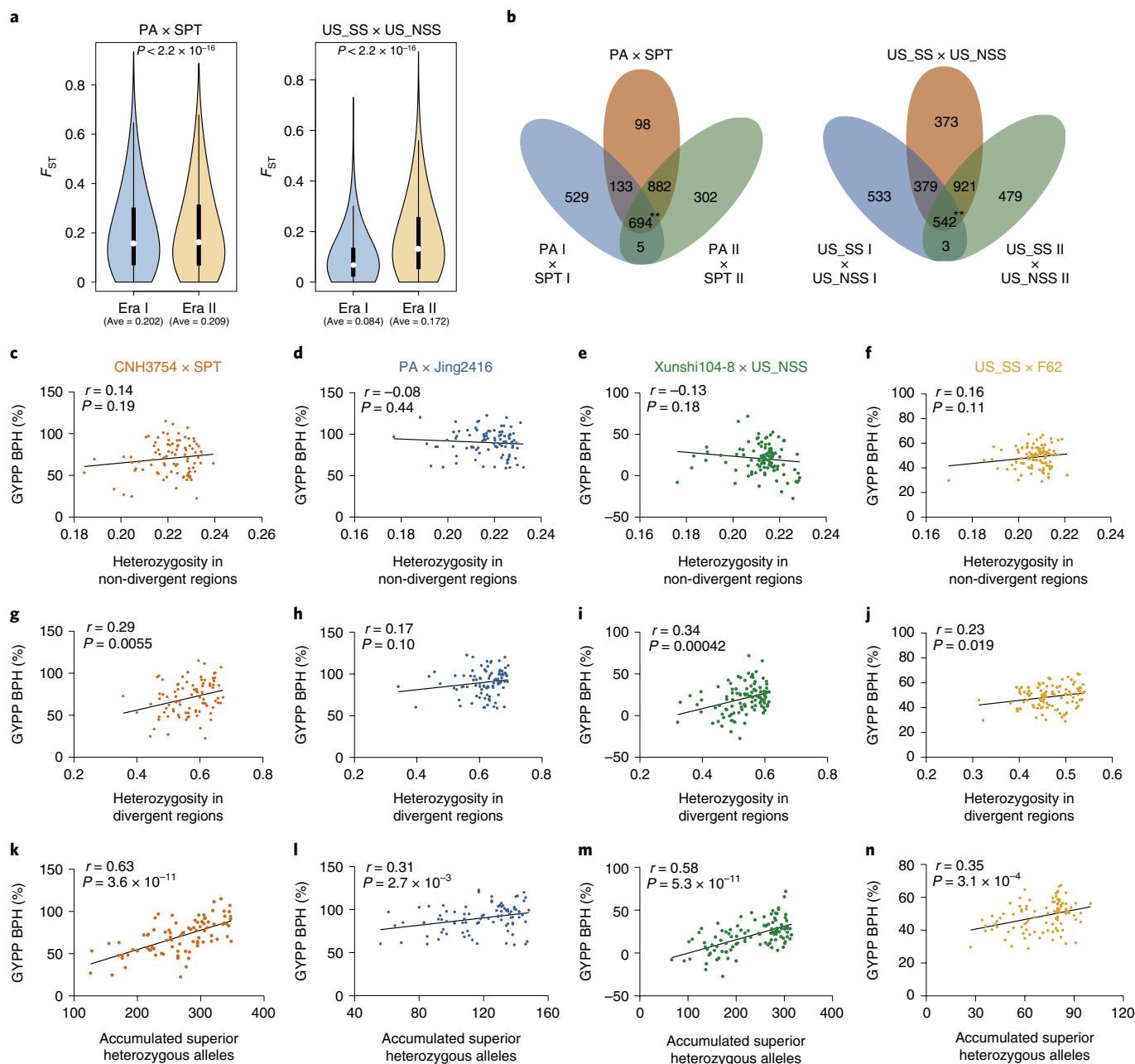
Fig. 6 | Genomic differentiation and effect of differentiated genes between the FHGs and MHGs on grain yield heterosis. a, Genome-wide changes of F_{ST} values between the female and male groups for the PA × SPT and US_SS × US_NSS heterotic patterns in different breeding eras. For the violin plots, the white dots represent the medians, the black box limits indicate the 25th and 75th percentiles, and the whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles. Statistical significance between Era I and Era II was determined using two-sided *t*-tests. Ave, average F_{ST} value in different breeding eras. **b**, The number of genes in the differentiated regions of two heterotic patterns across different breeding eras. Two asterisks (*P* < 0.01) indicate that the observed numbers are significantly greater than the expected numbers under permutation test conditions. **c–f**, Correlations between heterozygosity of non-synonymous SNPs in non-divergent regions and BPH of GYPP in four test-cross populations (CNH3754 × SPT (**c**), PA × Jing2416 (**d**), Xunshi104-8 × US_NSS (**e**) and US_SS × F62 (**f**)). **g–j**, Correlations between heterozygosity of non-synonymous SNPs located in genes continuously selected in the divergent regions and BPH of GYPP in four test-cross populations. A total of 2,563 and 2,136 non-synonymous SNPs contained in 478 and 375 continuously selected genes in the PA × SPT and US_SS × US_NSS patterns were used to calculate the heterozygosity for the test-cross populations of the PA × SPT and SS × NSS heterotic patterns, respectively. **k–n**, Correlations between the number of accumulated superior heterozygous alleles and BPH of GYPP in four test-cross populations. The four test-cross populations include 88 hybrids derived from CNH3754 (a PA inbred) × 88 SPT inbreds (**c**, **g** and **k**), 91 hybrids derived from 91 PA inbreds × Jing2416 (an SPT inbred) (**d**, **h** and **l**), 106 hybrids derived from Xunshi104-8 (an SS inbred) × 106 US_NSS inbreds (**e**, **i** and **m**) and 101 hybrids derived from 101 US_SS inbreds × F62 (an NSS inbred) (**f**, **j** and **n**). The Pearson correlation coefficient (*r*) and *P* value are presented in **c–n**.

Discussion

In this study, we conducted a comprehensive phenotypic assessment of 1,604 diverse inbred lines from different heterotic groups used in modern hybrid maize breeding, and we found several interesting features associated with the morphological changes of the FHGs and MHGs during modern maize breeding. First, we noted that the
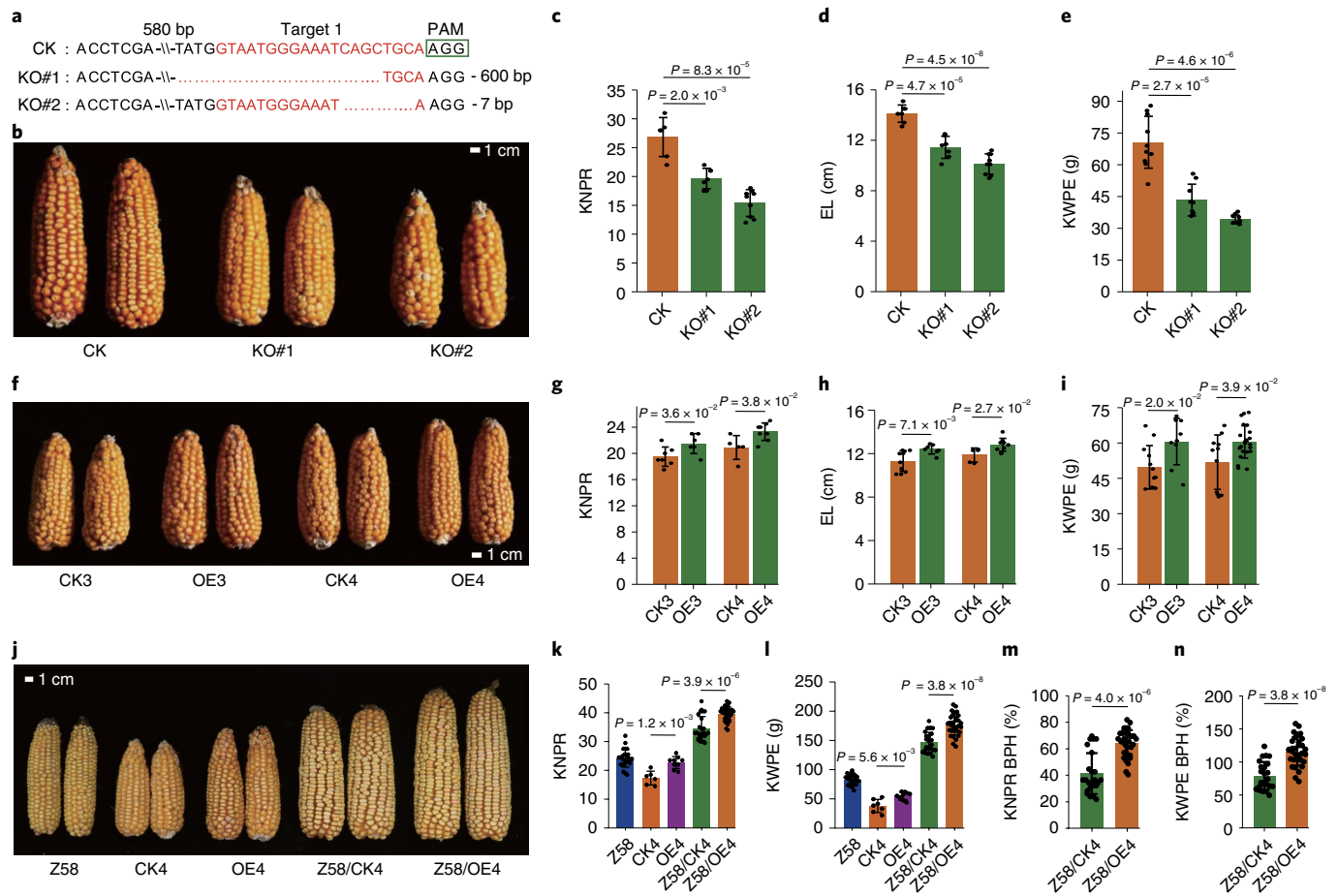
Fig. 7 | Validation of *ZmKOB1* as a differentiated gene between the PA and SPT heterotic groups. a, Knockout of *ZmKOB1* by CRISPR–Cas9 technology. **b–e**, Ear images (**b**) and statistics for KNPR (**c**), EL (**d**) and KWPE (**e**) of control plants (CK, null segregant plants with the wild-type *ZmKOB1* gene) and *Zmkob1* knockout lines (KO#1 and KO#2). The dots indicate biologically independent ears (n_{CK} = 6, $n_{KO\#1}$ = 7, $n_{KO\#2}$ = 9 in **c**; n_{CK} = 7, $n_{KO\#1}$ = 7, $n_{KO\#2}$ = 9 in **d**; and n_{CK} = 10, $n_{KO\#1}$ = 9, $n_{KO\#2}$ = 10 in **e**). **f–i**, Ear images (**f**) and statistics for KNPR (**g**), EL (**h**) and KWPE (**i**) of CK lines (CK3 and CK4) and *ZmKOB1* overexpression lines (OE3 and OE4). The dots indicate biologically independent ears (n_{CK3} = 7, n_{OE3} = 6, n_{CK4} = 5, n_{OE4} = 8 in **g**; n_{CK3} = 10, n_{OE3} = 7, n_{CK4} = 5, n_{OE4} = 10 in **h**; and n_{CK3} = 12, n_{OE3} = 9, n_{CK4} = 11, n_{OE4} = 23 in **i**). **j–n**, Ear phenotypes (**j**), yield-related data from Z58/OE4 and Z58/CK4 hybrids and their parents for KNPR (**k**) and KWPE (**l**), and comparison of BPH values between Z58/OE4 and Z58/CK4 hybrids for KNPR (**m**) and KWPE (**n**). The dots indicate biologically independent ears (from left to right, *n* = 26, 6, 9, 25 and 32 in **k**; *n* = 26, 7, 10, 24 and 32 in **l**; *n* = 25 and 32 in **m**; and *n* = 24 and 32 in **n**). The *P* values were estimated using two-sided *t*-tests. The data are presented as mean ± s.d.

FHGs in general have smaller tassels and higher yield-related traits (lower TBN and EP, and higher GYPP, KWPE, EW, KNPR, CW, KR, HKW, HKV and KL) than the MHGs across the two historical breeding eras (Fig. 2a–c). Second, we found that the MHGs seem to have experienced more extensive selection than the FHGs during modern hybrid maize breeding (18 traits in the MHGs and 10 traits in the FHGs were significantly changed from Era I to Era II). Third, consistent with previous reports that the increase in hybrid yield is accompanied by an increase in the yield of parental inbred lines[2,12], we found that the FHGs and MHGs both experienced convergent changes in flowering time, plant architecture and yield-related traits, including earlier DA and DS; shorter ASI; lower TBN, EH and EP; and higher GYPP, KWPE, EW, KR, KNPR, HKW, HKV, KL and KW (Fig. 2a,b). These changes are probably associated with adaptation for higher planting densities and greater grain yield in the advanced hybrids[2,3,15,46,47]. Notably, we found that the male parental lines might also have undergone more intensive (stronger) artificial selection for most convergently improved traits (such as KR, GYPP and KWPE) (Supplementary Table 5). Fourth, we observed divergent changes for ED, KRN and CW in the FHGs (decreases) and MHGs (increases) (Fig. 2c). The decrease of ED, KRN and CW in

the FHGs might be associated with the improved kernel dehydration rate in the female inbred lines and hybrids, which facilitates mechanical harvesting[48–50].

Our analyses also revealed a few interesting features associated with genomic changes that occurred during the improvement of maize heterotic groups. First, we found that on the basis of the detected GWAS signals, the accumulation of favourable alleles at the associated SNPs correlates well with the observed trait improvement in the FHGs and MHGs (Fig. 2d–f and Extended Data Fig. 3a–c). This observation suggests that these identified genes are associated with trait improvement of FHGs and MHGs during modern maize breeding. This notion is further verified by functional studies of *ZmEMF1L1* and *ZmKW10*, two convergently selected genes, in regulating flowering time/EH and KL/kernel weight, respectively, in the parental lines (Figs. 4 and 5). Second, we found that convergent selection occurred at both the gene and pathway levels (such as positive regulation of the defence response, response to water stimulus, starch binding, hormone binding and abscisic acid binding) in the FHGs and MHGs (Supplementary Table 10). This finding is consistent with the previous proposition that the new inbred lines and their hybrids are improved in their adaptation or tolerance

to various stresses (such as diseases, insects, drought, dense planting and low soil fertility)[2,3,51]. Third, we observed increased population differentiation during the genetic improvement of the FHGs and MHGs for both the PA×SPT and US_SS×US_NSS heterotic patterns, which probably contributed to promoting heterosis during modern hybrid maize breeding (Fig. 6). Notably, our results indicate that differential selection of different heterotic patterns largely targeted different genomic regions during modern maize breeding. This finding is consistent with the notion that different traits or different hybrids may utilize different genetic factors for heterosis[52,53]. Fourth, we found that while a small portion of the selected genes have been selected in both the FHGs and MHGs (20 of the 483 selected genes in the PA×SPT pattern and 82 of the 836 selected genes in the US_SS×US_NSS pattern), nearly half of the selected genes (226 of the 483 selected genes in the PA×SPT pattern and 464 of the 836 selected genes in the US_SS×US_NSS pattern) were selected in only one of the parental pools (Supplementary Table 16). These observations suggest that a significant percentage of the selected genes were fixed in one parental pool due to selection, and that complementation (dominance) of linked deleterious alleles by haplotypes in the other parental pool may explain their role in heterosis as previously proposed[54]. Fifth, we found that the increase of heterozygosity levels and the accumulation of favourable heterozygous alleles in the differentially selected regions/genes positively correlate with heterosis levels in their hybrids (Fig. 6 and Extended Data Fig. 7); this finding shares similarities with the observations in rice and *Arabidopsis*[55,56]. Finally, we validated the function of a differentiated gene (*ZmKOB1*) in regulating EL/EW in hybrids, through CRISPR–Cas9 knockout and overexpression studies (Fig. 7).

In summary, this study provides a genomic overview of the genetic improvement of the FHGs and MHGs during modern hybrid maize breeding and a rich resource of genetic variations. This should lay a foundation for further exploitation of the molecular basis of maize heterosis and facilitate the development of more effective and accurate genomic-informed maize breeding programmes.

Methods

Plant materials and phenotyping. The diversity panel used in this study comprised 1,604 maize inbred lines, including 1,227 from China, 344 from the United States and 33 from other countries. The 1,227 inbreds from China included a core collection of 242 inbreds selected from the 3,258 accessions preserved in the China National Crop Gene Bank[57], 345 parental lines of hybrids utilized in different breeding eras and 640 recently released lines. Because the Seed Law of the People's Republic of China including clauses related to Plant Variety Protection came into force in 2000, the Chinese maize breeding history can be roughly divided into two breeding eras: Era I (before 2000) and Era II (after 2000). The number of accessions from China in Era I is 408, and the number in Era II is 760. The 344 inbreds from the United States include 85 public lines (Era I) and 239 Ex-Plant Variety Protection lines (Era II) preserved in the China National Crop Gene Bank.

We selected 21 traits for phenotyping at five locations in 2018 and 2019. The five locations are Harbin (126° 63′ E, 45° 75′ N) in Heilongjiang Province, Shenyang (123° 38′ E, 41° 80′ N) in Liaoning Province, Beijing (116° 20′ E, 39° 56′ N), Zhengzhou (113° 65′ E, 34° 76′ N) in Henan Province and Urumqi (87° 30′ E, 43° 58′ N) in Xinjiang Autonomous Region. All inbred lines were grown in an experimental field with a randomized complete block design including two replicates in each environment. Each line was represented by 13 individual plants in a plot, with rows 3 m in length and 0.5–0.6 m between rows. In all the environments, 21 traits were measured, including 3 flowering-time traits, 5 plant-architecture-related traits and 13 yield-related traits (Supplementary Table 19). For flowering time, all plants in a plot were considered, while for plant architecture traits, three adjacent plants in the middle of the plot were measured and the average value was used. For ear-related traits, the measurements consisted of the average of five representative ears selected in a plot. For kernel traits, average size and weight measurements were made on 100 kernels from three different ears using an automatic seed testing machine.

A total of 386 test-crosses were produced by crossing inbred lines with the four common testers, including 88 hybrids derived from CNH3754 (a PA inbred)×88 SPT inbreds, 91 hybrids derived from 91 PA inbreds×Jing2416 (an SPT inbred), 106 hybrids derived from Xunshi104-8 (an SS inbred)×106 US_NSS inbreds and 101 hybrids derived from 101 US_SS inbreds×F62 (an NSS inbred). These test-crosses were all evaluated at two locations for two years (2018 and 2019) and were planted in the same field as the corresponding parental lines. In the field

trials at each location, these test-crosses were randomized in a complete block design with one replicate and were planted in two-row plots with a final plant density of 75,000 plants per ha. Three yield traits (GYPP, EW and KWPE) were measured for all environments, and the average values of each trait were used for subsequent analyses.

DNA isolation and genome sequencing. The genomic DNA was extracted with a total amount of 1.5 µg per sample and used as the input material for DNA isolation. Sequencing libraries were generated using the TruSeq Nano DNA HT Sample Preparation Kit (Illumina USA) following the manufacturer's recommendations, and index codes were added to attribute sequences to each sample. The libraries were prepared following these steps: the genomic DNA sample was fragmented by sonication to a size of ~350 base pairs (bp), and then the DNA fragments were end-polished, A-tailed and ligated with the full-length adapters for Illumina sequencing with further PCR amplification. The PCR products were then purified (AMPure XP bead system), and the libraries were analysed for size distribution by an Agilent 2100 Bioanalyzer and quantified using real-time PCR. Subsequently, we used the Illumina HiSeq X platform to generate 21.5 Tb raw sequences with 150-bp paired-end reads.

Sequence quality checking and filtering. To avoid reads with artificial bias (that is, low-quality paired reads, which primarily resulted from base-calling duplicates and adaptor contamination), we removed the following types of reads: (1) reads with ≥10% unidentified nucleotides (N); (2) reads with >10 nucleotides aligned to the adaptor, with ≤10% mismatches allowed; (3) reads with >50% bases having phred quality <5; and (4) putative PCR duplicates generated through PCR amplification in the library construction process (that is, read 1 and read 2 of two paired-end reads that were completely identical). We thus obtained 21.5 Tb (~7.5× coverage per individual) of high-quality paired-end reads, including 90.4% nucleotides with phred quality ≥Q30 (with an accuracy of 99.9%) (Supplementary Table 2).

Sequence alignment, variation calling and annotation. After sequence quality filtering, we first mapped the remaining sequences to the B73 reference genome (B73_V4, ftp://ftp.ensemblgenomes.org/pub/plants/release-37/fasta/zea_mays/dna)[58] using the software BWA (v.0.7.17-r1188)[59] with the command 'mem -t 10 -k 32 -M'. Second, we converted SAM format to BAM format using the package SAMtools (v.1.3)[60]. Third, we sorted the BAM files using the package Sambamba (v.0.6.8)[61]. Finally, we marked duplicate reads in the sorted BAM files using the command MarkDuplicates in the package picard (v.2.18.15, http://broadinstitute.github.io/picard). We then performed individual gVCF calling according to the best practices using the Genome Analysis Toolkit (GATK, v.4.1.2.0)[62] with the HaplotypeCaller-based method, followed by population SNP calling by merging all gVCFs with the commands GenomicsDBImport and GenotypeGVCFs. A total of 220,330,894 SNPs and 21,952,577 indels (small insertions and deletions <15 bp) were identified in the 1,604 maize inbred lines.

To obtain credible population SNP sets, we performed a screening process as follows:

(1) For filtering SNPs, the hard filter command VariantFiltration was applied to exclude potential false-positive variant calls with the parameter '–filterExpression "QD < 2.0 || MQ < 40.0 || FS > 60.0 || SOR > 3.0 || MQRankSum < −12.5 || ReadPosRankSum < −8.0"'.
(2) Variants were filtered out when the proportion of samples within the population lacking the variant was ≥20%, the minor allele frequency was ≤0.01 and the heterozygosity rate was ≥15%. After those steps, we obtained 18,169,560 biallelic SNPs that were used for subsequent analyses. To validate SNP quality, we first compared our data with the data from Hapmap3 for 19 lines with sequencing depth >5.0×[63] and found that the SNP concordance rates reached 94.5–97.9% (Supplementary Table 20). Second, we compared our data with the Illumina 50K SNP chip genotypic data for an association panel of 41 lines of 282 inbreds[64] and found that the SNP concordance rates reached 96.7–98.8% (Supplementary Table 21).

Genomic variant annotation was performed according to the B73 reference genome using the package ANNOVAR[65]. On the basis of the genome annotation, genomic variants were categorized as being in exonic regions, intronic regions, splice sites (within 2 bp of a splicing junction), upstream and downstream regions (within a 1-kb region upstream or downstream of the transcription start site), and intergenic regions. The functional consequences of the variants in coding regions were further grouped into synonymous, non-synonymous, stop-gain and stop-loss.

Clustering and population structure analyses. Population genetic structure was examined using the Bayesian clustering program fastStructure (v.1.0)[66], and different levels of K (K = 2 to 15) were calculated to determine the optimal number of subpopulations on the basis of the bestK module. In the end, K = 11 is a reasonable number for our group division (Extended Data Fig. 1b). To verify the rationality of the 11 subgroups, a PCA was conducted using the software GCTA[67] with all 18,169,560 SNPs. We first obtained the genetic relationship matrix with the parameter '–make-grm'. Then, the top three principal components were estimated with the parameter '–pca3'. Finally, we also estimated an individual-based

neighbour-joining tree on the basis of the *p*-distance using the software TreeBest (v.1.9.2)[68] with 1,000 bootstrap replications.

Analysis of LD. To estimate the LD for the 1,604 lines, we calculated the squared correlation coefficient (r^2) between pairwise SNPs using the software PopLDdecay[69]. The program parameters were set as '-MaxDist 100 -MAF 0.01 -Miss 0.2' to calculate the average r^2 between two SNPs in 100-kb windows. The overall LD decay distance in the 1,604 lines was 76 bp ($r^2 = 0.2$) and 2.5 kb (half of max r^2).

Identification of selective sweeps for breeding improvement. To identify potential selective signals within each heterotic group during modern maize breeding (PA II versus PA I, US_SS II versus US_SS I, PB II versus PB I, SPT II versus SPT I, CN_NSS II versus CN_NSS I and US_NSS II versus US_NSS I), we used the XP-CLR score[70] to scan for selective sweep regions (−w1 0.02 800 10000 1−p0 0.95). The windows with the top 5% of XP-CLR scores were considered as candidate sweeps during breeding improvement, and genes located in or overlapping with selective sweeps were identified as potential candidate genes. To facilitate statistical results, the neighbouring selective sweeps were then merged into one sweep if the distance between two selective sweeps was less than 20 kb. The final merged regions were considered as selective sweeps within each heterotic group. To obtain the total number of selective sweeps across six heterotic groups, we considered selective sweeps from different groups with at least 10 kb of overlap as the same.

Identification of differentiated regions. We estimated F_{ST} and AFD in 20-kb sliding windows with a step size of 10 kb to quantify genomic differentiation between the female pool and male pool using VCFtools (v.0.1.15)[71]. Sliding windows with the top 5% highest F_{ST} or AFD values were selected initially. Outlier windows less than 20 kb were merged. Differentiated regions were identified as those in the top 5% of both statistics.

Beagle (v.4.1)[72] with the default parameters was used to infer missing genotypes in the 100 kb upstream and downstream of gene regions.

Estimating breeding value. A BLUP value of each trait across all environmental trials was obtained for each inbred line by a linear mixed model in R (v.3.6.1) (http://www.r-project.org/) with the lme4 package[73]. The formula is as follows:

$$Y = \mu + \text{Line} + \text{Env} + (\text{Line} \times \text{Env}) + \text{Rep}(\text{Env}) + \varepsilon$$

where Y, μ, Line and Env represent phenotype, mean, inbred line effects and environmental effects, respectively; Rep means different replications; and ε represents the pooled error. Line × Env is used to display the interaction between inbred line and environment, and Rep(Env) shows the nested effect of replication within environment. These items (Line, Env, Line × Env and Rep(Env)) are set to random effects.

GWAS analysis. A GWAS was performed using the software EMMAX[74] and all 18,169,560 high-quality SNPs. The kinship matrix of pairwise genetic similarities, which were derived from the simple matching coefficients, was used as the variance–covariance matrix of the random effects and was also calculated by EMMAX. We conducted GWAS using both the BLUP and single trial values for the 21 traits. To determine the genome-wide threshold, we conducted permutation tests by randomly shuffling the phenotypes for three representative traits (GYPP, PH and DS). We then applied GWAS on the permuted phenotypes by using the same model that was used for the real observed phenotypes. The most significant *P* value across the whole genome was recorded. This random process was repeated 100 times for each trait. The distribution of the most significant *P* values across the 100 replicates was used to determine the threshold, which was the *P* value corresponding to a 5% chance of a type I error. The three traits had very similar thresholds ($-\log_{10}(P) = 6.1–6.5$). We then used a threshold of $P < 1.0 \times 10^{-6}$ to identify significant association signals for subsequent analysis.

Significant signals were determined with the following two criteria: (1) SNP signals were significantly associated with the BLUP values, or (2) SNP signals were significantly and repeatedly associated with values from at least two environmental trials. The local LD blocks with $r^2 \geq 0.7$ around peak SNPs (above the threshold) were defined as the candidate associated regions. Genes within the candidate associated regions were selected as the candidate genes for the GWAS associations. On the basis of the associated SNPs, the allele types (reference allele or alternative allele) that confer better agronomic performance (for example, earlier DA and DS; shorter ASI; lower EH, EP and TBN; and higher GYPP, KWPE, EW, KNPR, KR, KRN, ED, CW, HKW, HKV, KL and KW) were deemed the favourable alleles.

GO analysis. GO annotation terms for maize genes (B73 v.4.0 assembly) were downloaded from the Ensembl Plants Genes database (https://plants.ensembl.org/biomart/martview/). GO analysis was performed with agriGO (v.2.0)[75]. Enrichment significance was analysed with the Fisher test. Enrichment results with more than five annotations and a Bonferroni false discovery rate of <0.05 were plotted with the R package ClusterProfiler (v.3.10.0)[76].

Gene expression analysis. RNA was extracted using the Plant Total RNA Isolation Kit (Genebetter, catalogue no. R318-50). First-strand complementary DNA

was synthesized using the Uni One-Step gDNA Removal and cDNA Synthesis SuperMix Kit (TranScript, catalogue no. AU311). Quantitative real-time PCR was performed using the PowerUp SYBR Green Master Mix (Applied Biosystems, catalogue no. A25742) and a QuanStudio 3 Real Time PCR System cycler (Applied Biosystems). The maize *GAPDH* gene (Zm00001d049641) was used as the internal control. The relative expression of the gene was calculated using the $2^{-\Delta\Delta Ct}$ method. The primers used for quantitative real-time PCR are listed in Supplementary Table 22.

Gene cloning and plant transformation. The gene editing and overexpression vectors were constructed according to a previously described protocol[77]. For gene editing of *ZmEMF1L1*, *ZmKW10* and *ZmKOB1*, all single guide RNA (sgRNA) constructs were designed and introduced into sgRNAs expression cassettes by overlapping PCR. The three sgRNA expression cassettes of both *ZmEMF1L1* and *ZmKW10* and two sgRNA expression cassettes of *ZmKOB1* were then integrated into the pCPB–ZmUbi::hSpCas9 vectors. In addition, to construct the *ZmKOB1* overexpression vector, the full-length coding sequence of *ZmKOB1* was amplified from B73 and inserted into a CPB–proZmUbi::GFP vector to generate the proZmUbi::*ZmKOB1*–GFP construct. The gene-editing constructs of both *ZmEMF1L1* and *ZmKW10* were introduced into the *Agrobacterium* strain EHA105 and transformed into the immature embryo of the maize inbred line B104 through *Agrobacterium*-mediated transformation. Both of the gene-editing and overexpression constructs of *ZmKOB1* were transferred to the *Agrobacterium* strain EHA105 and transformed into the immature embryo of the maize inbred line ZC01 through *Agrobacterium*-mediated transformation.

The *Zmemf1L1* knockout mutants of T$_2$ generation and wild-type plants were planted at Beijing (116° 20′ E, 39° 56′ N), China, in spring 2021 using a randomized block design with three replicates. Each plot consisted of 26 individuals grown in two rows 3 m in length, with a spacing of 0.25 m between plants and 0.6 m between rows. PH was measured from the soil level to the tip of the main inflorescence, and EH was measured from the soil level to the node of the primary ear. DA was measured from sowing to anthesis. The *Zmkw10* knockout mutants of the T$_2$ generation were planted in winter 2020 at Sanya (109° 48′ E, 18° 09′ N), Hainan Province, China. After harvesting and air-drying, the representative ears were used to measure 10 KL, 10 KW and HKW. Phenotypic evaluation of the CK line (null segregant plants with the wild-type *ZmKOB1* gene), the *Zmkob1* knockout mutants and the overexpression lines of the T$_3$ generation was performed in fall 2020 at Langfang (116° 38′ E, 39° 28′ N), Hebei Province, China. After harvesting, the ears were air-dried, and KWPE, KNPR and EL were measured. In addition, the Z58/OE4 and Z58/CK4 hybrids were generated by crossing the inbred line Zheng58 (maternal parent) with *ZmKOB1–OE4* and the control line (CK4), respectively. For the field experiments, Zheng58, CK4, OE4 and F$_1$ hybrids were planted in Langfang (116° 38′ E, 39° 28′ N), Hebei Province, China, during the spring of 2021. After harvesting, air-dried ears were used to measure KNPR and KWPE. The primers used above are listed in Supplementary Table 22.

Heterosis analysis of maize hybrids. BPH describes the degree of phenotypic difference between a hybrid and its better parent. BPH was calculated using the following formula: $\text{BPH} = (F_1 - BP)/BP$, where F_1 is the phenotypic value of the hybrid, and BP is the phenotypic value of the better-performing parent. The effects of heterozygous alleles and homozygous alleles were analysed for non-synonymous SNPs of continuously differentiated genes for GYPP, EW and KWPE. The non-synonymous SNPs with significantly different effects between the heterozygous state and homozygous state (Wilcoxon test) were identified as superior alleles. For each non-synonymous SNP, the heterozygous allele with better yield performance (higher GYPP, EW and KWPE) was deemed the superior heterozygous allele.

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

Data availability

The raw sequencing data were deposited in the Genome Sequence Archive (https://bigd.big.ac.cn/gsa) under the accession code PRJCA009749 (https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA009749). The phenotype dataset reported here is available from the corresponding authors upon request. Source data are provided with this paper.

Code availability

The custom codes used in this study are available at https://github.com/jasongit0311/maize_for_Li.

References

1. *Food and Agriculture Organization of the United Nations Agriculture Databases* (FAO, 2019); http://www.fao.org/statistics/databases/en/

2. Duvick, D. N. The contribution of breeding to yield advances in maize (*Zea mays* L.). *Adv. Agron.* **86**, 83–145 (2005).

3. Duvick, D. N. Genetic progress in yield of United States maize (*Zea mays* L.). *Maydica* **50**, 193–202 (2005).

4. Mansfield, B. D. & Mumm, R. H. Survey of plant density tolerance in U.S. maize germplasm. *Crop Sci.* **54**, 157–173 (2014).

5. Andorf, C. et al. Technological advances in maize breeding: past, present and future. *Theor. Appl. Genet.* **132**, 817–849 (2019).

6. Tracy, W. F. & Chandler, M. A. in *Plant Breeding: The Arnel R. Hallauer International Symposium* Ch. 16 (eds Lamkey, K. R. & Lee, M.) (Blackwell, 2006).

7. Mikel, M. A. & Dudley, J. W. Evolution of North American dent corn from public to proprietary germplasm. *Crop Sci.* **46**, 1193–1205 (2006).

8. Lu, Y. L. et al. Molecular characterization of global maize breeding germplasm based on genome-wide single nucleotide polymorphisms. *Theor. Appl. Genet.* **120**, 93–115 (2009).

9. Melchinger, A. E. & Gumber, R. K. in *Concepts and Breeding of Heterosis in Crop Plants* pp. 29–44 (eds Larnkey, K. R. & Staub, J. E.) (Crop Science Society of America, 1998).

10. Reif, J. C., Hallauer, A. R. & Melchinger, A. E. Heterosis and heterotic patterns in maize. *Maydica* **50**, 215–223 (2005).

11. Lauer, S. et al. Morphological changes in parental lines of pioneer brand maize hybrids in the U.S. Central Corn Belt. *Crop Sci.* **52**, 1033–1043 (2012).

12. Li, Y. X. et al. Contributions of parental inbreds and heterosis to morphology and yield of single-cross maize hybrids in China. *Crop Sci.* **54**, 76–88 (2014).

13. Gage, J. L., White, M. R., Edwards, J. W., Kaeppler, S. & de Leon, N. Selection signatures underlying dramatic male inflorescence transformation during modern hybrid maize breeding. *Genetics* **210**, 1125–1138 (2018).

14. Wang, B. B. et al. Genome-wide selection and genetic improvement during modern maize breeding. *Nat. Genet.* **52**, 565–571 (2020).

15. Duvick, D. N., Smith, J. S. C. & Cooper, M. in *Plant Breeding Reviews, Part 2: Long-Term Selection—Crops, Animals, and Bacteria* pp. 109–151 (ed. Janick, J.) (John Wiley & Sons, 2004).

16. van Heerwaarden, J., Hufford, M. B. & Ross-Ibarra, J. Historical genomics of North American maize. *Proc. Natl Acad. Sci. USA* **109**, 12420–12425 (2012).

17. Jiao, Y. P. et al. Genome-wide genetic changes during modern breeding of maize. *Nat. Genet.* **44**, 812–815 (2012).

18. Lu, H. & Bernardo, R. Molecular marker diversity among current and historical maize inbreds. *Theor. Appl. Genet.* **103**, 613–617 (2001).

19. Unterseer, S. et al. A comprehensive study of the genomic differentiation between temperate Dent and Flint maize. *Genome Biol.* **17**, 137 (2016).

20. Wu, X. et al. Analysis of genetic differentiation and genomic variation to reveal potential regions of importance during maize improvement. *BMC Plant Biol.* **15**, 256 (2015).

21. Li, C. H. et al. The HuangZaoSi maize genome provides insights into genomic variation and improvement history of maize. *Mol. Plant* **12**, 402–409 (2019).

22. Mikel, M. A. Genetic composition of contemporary US commercial dent corn germplasm. *Crop Sci.* **51**, 592–599 (2011).

23. Zhang, R. Y. et al. Patterns of genomic variation in Chinese maize inbred lines and implications for genetic improvement. *Theor. Appl. Genet.* **131**, 1207–1221 (2018).

24. Salvi, S. et al. Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proc. Natl Acad. Sci. USA* **104**, 11376–11381 (2007).

25. Strable, J. & Vollbrecht, E. Maize *YABBY* genes *drooping leaf1* and *drooping leaf2* regulate floret development and floral meristem determinacy. *Development* **146**, dev171181 (2019).

26. Provencher, L. M., Miao, L., Sinha, N. & Lucas, W. J. *Sucrose export defective1* encodes a novel protein implicated in chloroplast-to-nucleus signaling. *Plant Cell* **13**, 1127–1141 (2001).

27. Galli, M. et al. Auxin signaling modules regulate maize inflorescence architecture. *Proc. Natl Acad. Sci. USA* **112**, 13372–13377 (2015).

28. Wallace, J. G. et al. Association mapping across numerous traits reveals patterns of functional variation in maize. *PLoS Genet.* **10**, e1004845 (2014).

29. Yang, Q. et al. CACTA-like transposable element in *ZmCCT* attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize. *Proc. Natl Acad. Sci. USA* **110**, 16969–16974 (2013).

30. Huang, C. et al. *ZmCCT9* enhances maize adaptation to higher latitudes. *Proc. Natl Acad. Sci. USA* **115**, E334–E341 (2018).

31. Jia, H. T. et al. A serine/threonine protein kinase encoding gene *KERNEL NUMBER PER ROW6* regulates maize grain yield. *Nat. Commun.* **11**, 998 (2020).

32. Aubert, D. et al. EMF1, a novel protein involved in the control of shoot architecture and flowering in *Arabidopsis*. *Plant Cell* **13**, 1865–1875 (2001).

33. Melchinger, A. E. in *The Genetics and Exploitation of Heterosis in Crops* Ch. 10 (eds Coors, J. G. & Pandey, S.) (American Society of Agronomy, Crop Science Society of America and Soil Science Society of America, 1999).

34. Reif, J. C. et al. Genetic distance based on simple sequence repeats and heterosis in tropical maize populations. *Crop Sci.* **43**, 1275–1282 (2003).

35. Reif, J. C., Gumpert, F. M., Fischer, S. & Melchinger, A. E. Impact of interpopulation divergence on additive and dominance variance in hybrid populations. *Genetics* **176**, 1931–1934 (2007).

36. Technow, F. et al. Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. *Genetics* **197**, 1343–1355 (2014).

37. Chen, L. et al. Portrait of a genus: the genetic diversity of *Zea*. Preprint at *bioRxiv* https://doi.org/10.1101/2021.04.07.438828 (2021).

38. Jiang, F. K. et al. Mutations in an *AP2* transcription factor-like gene affect internode length and leaf shape in maize. *PLoS ONE* **7**, e37040 (2012).

39. Sawers, R. J. H. et al. The *Elm1* (*ZmHy2*) gene of maize encodes a phytochromobilin synthase. *Plant Physiol.* **136**, 2771–2781 (2004).

40. Danilevskaya, O. N. et al. Involvement of the MADS-box gene *ZMM4* in floral induction and inflorescence development in maize. *Plant Physiol.* **147**, 2054–2069 (2008).

41. Xia, H., Yandeau-Nelson, M., Thompson, D. B. & Guiltinan, M. J. Deficiency of maize starch-branching enzyme I results in altered starch fine structure, decreased digestibility and reduced coleoptile growth during germination. *BMC Plant Biol.* **11**, 95 (2011).

42. Bai, L., Kim, E. H., DellaPenna, D. & Brutnell, T. P. Novel lycopene epsilon cyclase activities in maize revealed through perturbation of carotenoid biosynthesis. *Plant J.* **59**, 588–599 (2009).

43. Wang, X., Jing, Y. J., Zhang, B. C., Zhou, Y. H. & Lin, R. C. Glycosyltransferase-like protein ABI8/ELD1/KOB1 promotes *Arabidopsis* hypocotyl elongation through regulating cellulose biosynthesis. *Plant Cell Environ.* **38**, 411–422 (2015).

44. Tognetti, V. B. et al. Perturbation of indole-3-butyric acid homeostasis by the UDP-glucosyltransferase UGT74E2 modulates *Arabidopsis* architecture and water stress tolerance. *Plant Cell* **22**, 2660–2679 (2010).

45. Dong, N. Q. et al. UDP-glucosyltransferase regulates grain size and abiotic stress tolerance associated with metabolic flux redirection in rice. *Nat. Commun.* **11**, 2629 (2020).

46. Duvick, D. N. in *Developing Drought and Low N-Tolerant Maize* pp. 332-335 (eds Edmeades, G. O. et al.) (CIMMYT, 1997).

47. Wang, T. Y. et al. Changes in yield and yield components of single-cross maize hybrids released in China between 1964 and 2001. *Crop Sci.* **51**, 512–525 (2011).

48. Purdy, J. L. & Crane, P. L. Inheritance of drying rate in "Mature" corn (*Zea mays* L.). *Crop Sci.* **7**, 294–297 (1967).

49. Zhou, G. F. et al. Genome-wide association study of kernel moisture content at harvest stage in maize. *Breed. Sci.* **68**, 622–628 (2018).

50. Wang, K. R. & Li, S. K. Analysis of influencing factors on kernel dehydration rate of maize hybrids. *Sci. Agric. Sin.* **50**, 2027–2035 (2017).

51. Troyer, A. F. Adaptedness and heterosis in corn and mule hybrids. *Crop Sci.* **46**, 528–543 (2006).

52. Schnable, P. S. & Springer, N. M. Progress toward understanding heterosis in crop plants. *Annu. Rev. Plant Biol.* **64**, 71–88 (2013).

53. Springer, N. M. & Stupar, R. M. Allelic variation and heterosis in maize: how do two halves make more than a whole? *Genome Res.* **17**, 264–275 (2007).

54. Gerke, J. P. et al. The genomic impacts of drift and selection for hybrid performance in maize. *Genetics* **201**, 1201–1211 (2015).

55. Huang, X. H. et al. Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. *Nat. Commun.* **6**, 6258 (2015).

56. Yang, M. et al. Genomic architecture of biomass heterosis in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **114**, 8101–8106 (2017).

57. Li, Y., Shi, Y. S., Cao, Y. S. & Wang, T. Y. Establishment of a core collection for maize germplasm preserved in Chinese National GeneBank using geographic distribution and characterization data. *Genet. Resour. Crop Evol.* **51**, 845–852 (2004).

58. Jiao, Y. P. et al. Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524–527 (2017).

59. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

60. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

61. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).

62. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

63. Bukowski, R. et al. Construction of the third-generation *Zea mays* haplotype map. *GigaScience* **7**, gix134 (2018).

64. Cook, J. P. et al. Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. *Plant Physiol.* **158**, 824–834 (2012).

65. Wang, K., Li, M. Y. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

66. Raj, A., Stephens, M. & Pritchard, J. K. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197**, 573–589 (2014).

67. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

68. Vilella, A. J. et al. Ensemblcompara genetrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19**, 327–335 (2009).

69. Zhang, C., Dong, S. S., Xu, J. Y., He, W. M. & Yang, T. L. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786–1788 (2019).

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

71. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

72. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).

73. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).

74. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).

75. Tian, T. et al. agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res.* **45**, W122–W129 (2017).

76. Yu, G., Wang, L. G., Han, Y. & He, Q.Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

77. Li, Q. Q. et al. CRISPR/Cas9-mediated knockout and overexpression studies reveal a role of maize phytochrome C in regulating flowering time and plant height. *Plant Biotechnol. J.* **18**, 2520–2532 (2020).

Acknowledgements

Author contributions

T.W., Yu Li, H.W. and C. Li conceived and designed the experiments. T.W., Yu Li, C. Li, Yongxiang Li, Y. Shi and Y. Song participated in the germplasm preparation. C. Li, X.J. and C.J. performed bioinformatics analyses of the data. H.G., Yaoyao Li and C. Li carried out the validation of gene function and gene expression analysis. B.W., H.G., Yaoyao Li, Yongxiang Li and X.L. participated in the data analysis. C. Li, D.Z., C. Liu, X.X., H.Z., Y.W., J.L., P.Z., G.H., G.L., S.L., D.S., X.W., Y. Shi and Y. Song performed the phenotypic measurements. C. Li, H.W., J.R.-I., Yu Li and T.W. analysed the data and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41477-022-01190-2.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41477-022-01190-2.
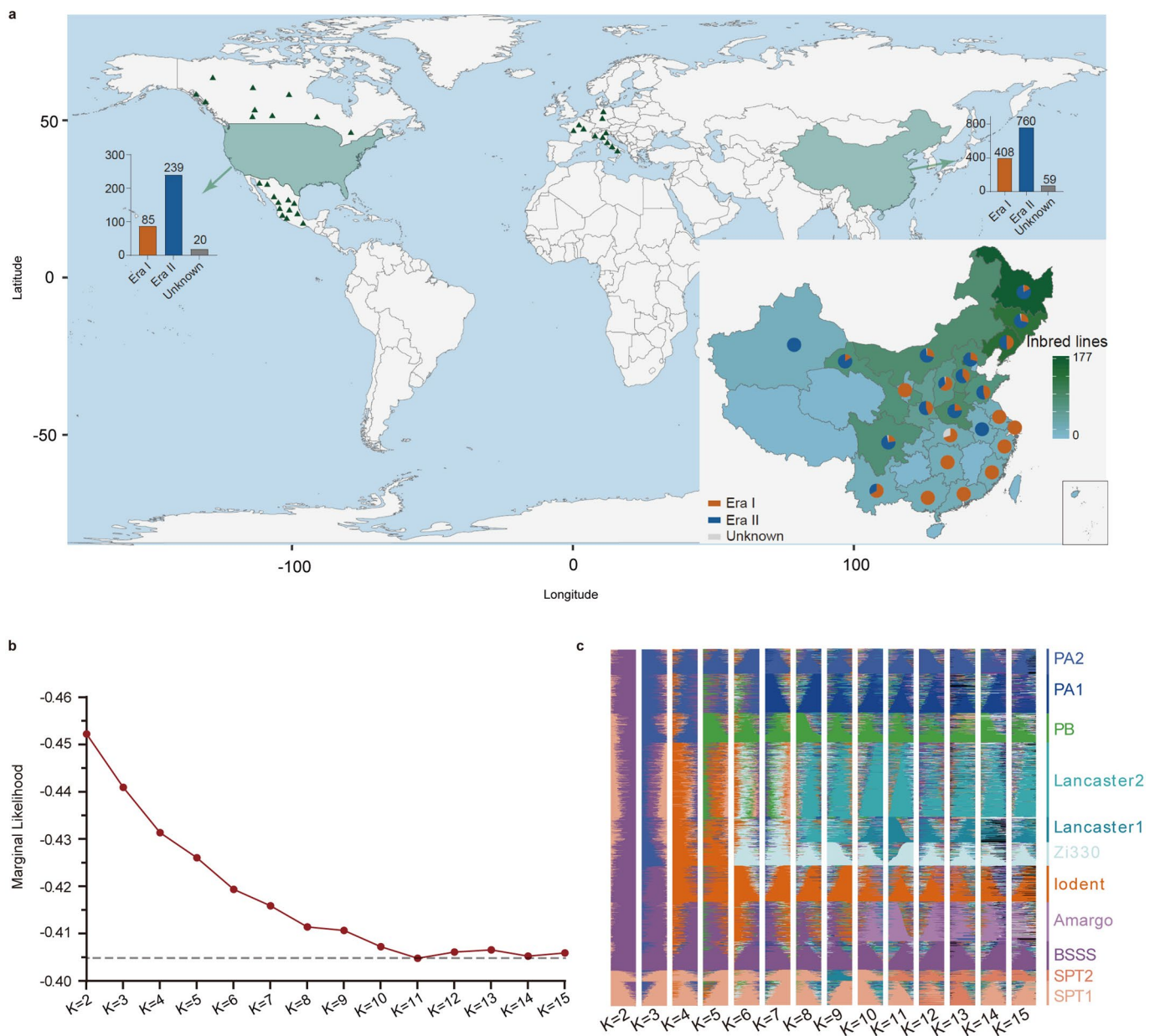
Correspondence and requests for materials should be addressed to Chengzhi Jiao, Jeffrey Ross-Ibarra, Yu Li, Tianyu Wang or Haiyang Wang.

Peer review information *Nature Plants* thanks Jianming Yu, Klaus Mayer, Xuehui Huang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
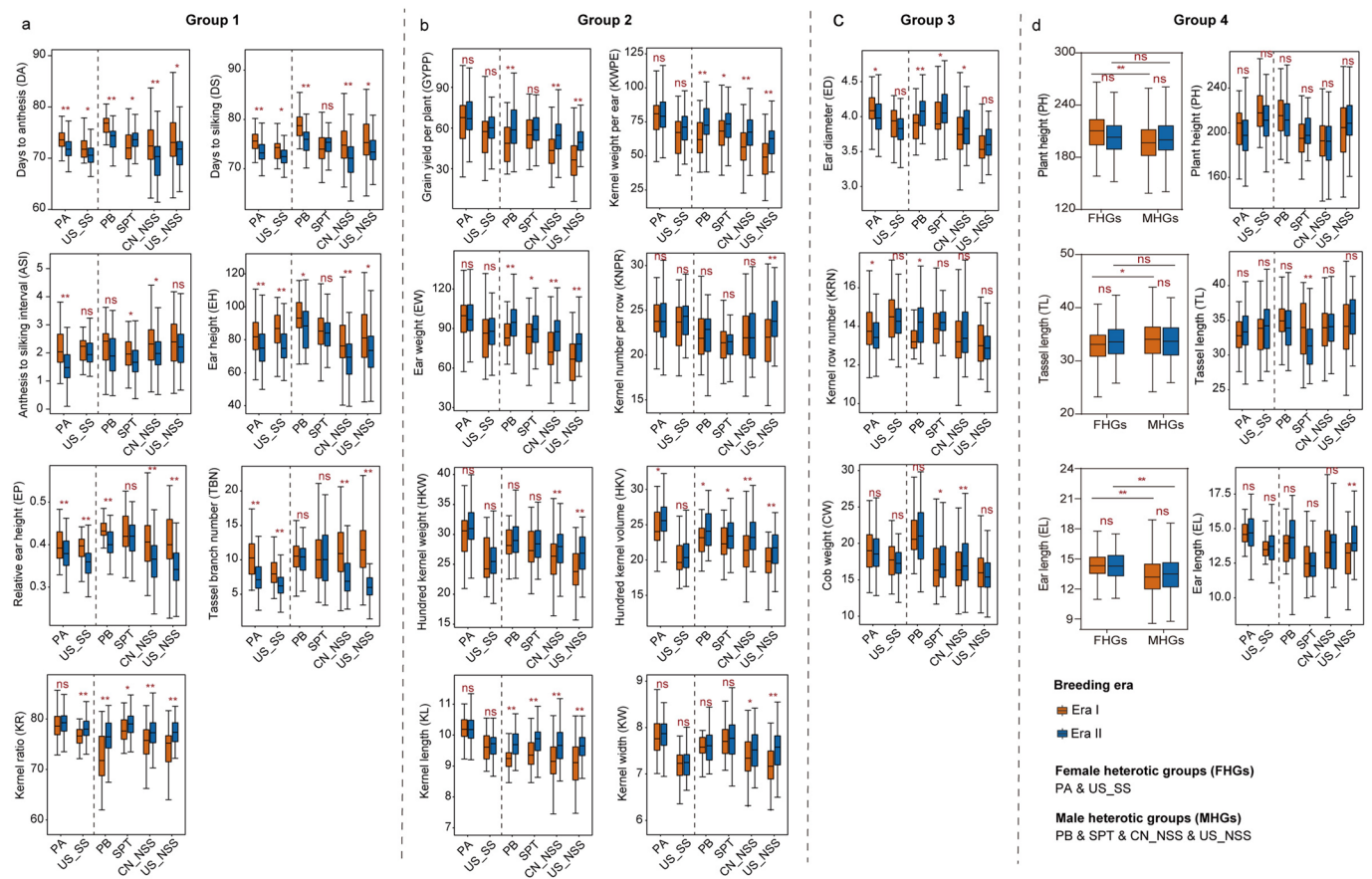
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
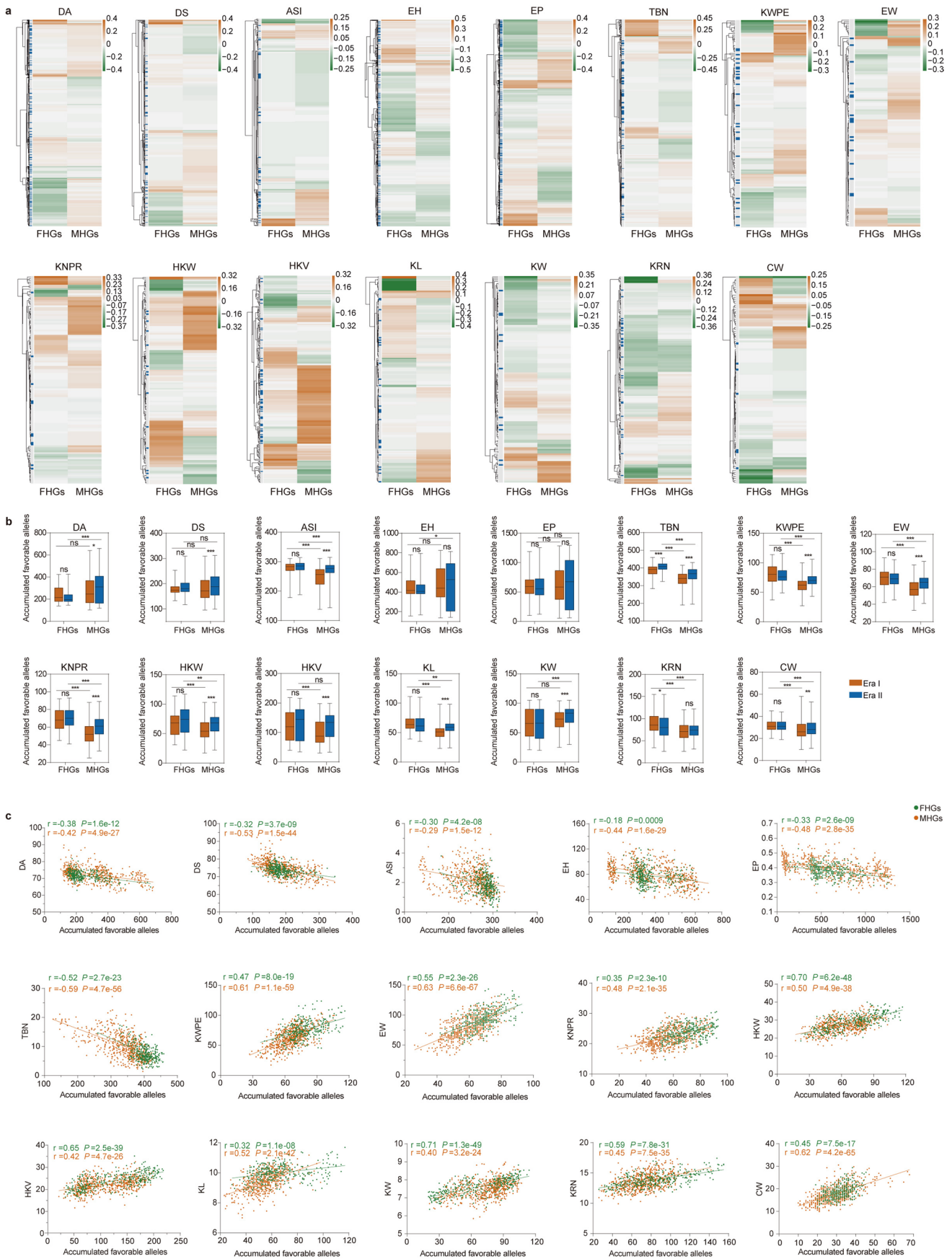
Extended Data Fig. 1 | Geographic distribution and population structure analyses of the 1,604 inbred lines. a, Geographic distribution of all resequenced inbreds according to their origin. Bar graphs show the number of inbreds released in Era I and Era II in the US and China. The orange and blue pie charts show the proportions of inbreds released in Era I and Era II, respectively. The map was drawn using the R ggmap package (http://maps.stamen.com/, map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL). **b**, The marginal likelihood begins to plateau at a $K=11$, which was considered as the optimal number of subgroups based on the fastStructure analysis. **c**, Population structure of the panel analyzed from $K=2$ to 15. $K=11$ clearly divided the panel into eleven subgroups, including PA1, PA2, PB, Lancaster1, Lancaster2, Zi330, Iodent, Amargo, BSSS, SPT1 and SPT2.
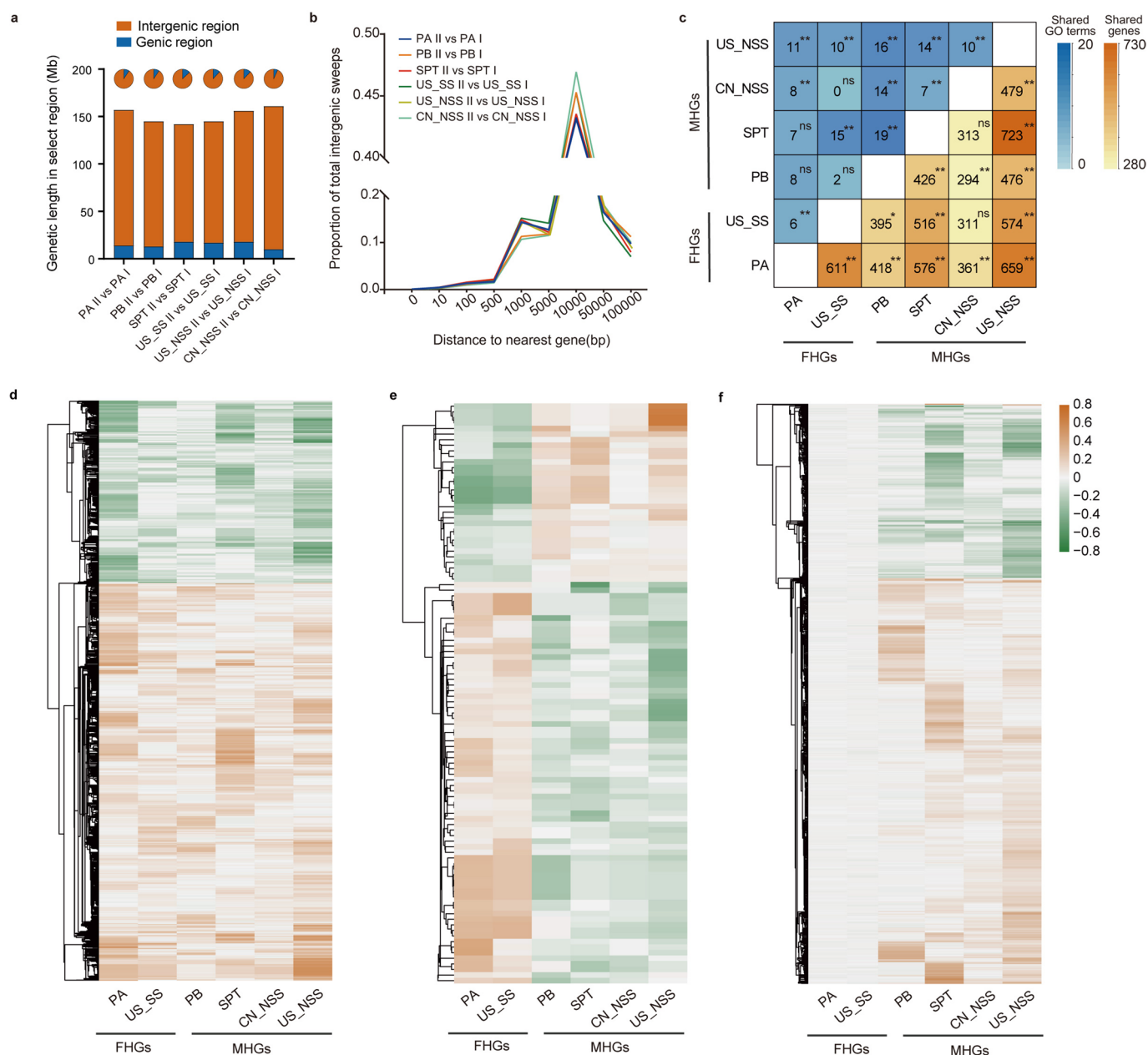
Extended Data Fig. 2 | Phenotypic changes in the FHGs and MHGs across different breeding eras. a-c, phenotypic changes in the two female groups (PA and US_SS) and four male groups (PB, SPT, CN_NSS and US_NSS) from Era I to Era II. Group 1 (**a**), Group 2 (**b**) and Group 3 (**c**) include seven, eight and three traits, respectively. The P values of Era I versus Era II in six groups using two-sided Wilcoxon test (left to right, as plotted) are: DA, $P=6.36\times10^{-6}$, 3.73×10^{-2}, 1.34×10^{-5}, 2.07×10^{-2}, 8.83×10^{-6} and 2.09×10^{-2}, respectively; DS, $P=7.40\times10^{-8}$, 1.27×10^{-2}, 5.83×10^{-5}, 9.61×10^{-2}, 1.00×10^{-6} and 1.83×10^{-2}, respectively; ASI, $P=4.14\times10^{-7}$, 1.58×10^{-1}, 1.29×10^{-1}, 4.64×10^{-2}, 1.05×10^{-2} and 2.64×10^{-1}, respectively; EH, $P=5.35\times10^{-3}$, 3.14×10^{-4}, 3.27×10^{-2}, 6.25×10^{-1}, 7.50×10^{-6} and 1.81×10^{-2}, respectively; EP, $P=3.94\times10^{-3}$, 2.93×10^{-4}, 1.39×10^{-3}, 3.71×10^{-1}, 4.12×10^{-8} and 1.30×10^{-6}, respectively; TBN, $P=1.07\times10^{-7}$, 1.26×10^{-4}, 6.45×10^{-1}, 9.31×10^{-1}, 2.95×10^{-12} and 2.62×10^{-10}, respectively; KR, $P=3.55\times10^{-1}$, 3.55×10^{-3}, 2.58×10^{-3}, 1.55×10^{-2}, 2.21×10^{-4} and 9.67×10^{-5}, respectively; GYPP, $P=3.67\times10^{-1}$, 1.68×10^{-1}, 9.67×10^{-3}, 2.13×10^{-1}, 3.26×10^{-8} and 3.12×10^{-6}, respectively; KWPE, $P=7.93\times10^{-1}$, 2.56×10^{-1}, 1.76×10^{-3}, 2.09×10^{-2}, 1.46×10^{-7} and 2.15×10^{-5}, respectively; EW, $P=9.48\times10^{-1}$, 4.39×10^{-1}, 4.85×10^{-3}, 3.47×10^{-2}, 1.42×10^{-6} and 4.72×10^{-4}, respectively; KNPR, $P=5.77\times10^{-1}$, 2.05×10^{-1}, 3.37×10^{-1}, 9.58×10^{-1}, 6.71×10^{-1} and 1.66×10^{-4}, respectively; HKW, $P=6.13\times10^{-2}$, 4.09×10^{-1}, 4.53×10^{-1}, 2.59×10^{-1}, 6.90×10^{-4} and 4.29×10^{-4}, respectively; HKV, $P=2.15\times10^{-2}$, 5.56×10^{-1}, 2.78×10^{-2}, 3.84×10^{-2}, 8.49×10^{-7} and 9.94×10^{-4}, respectively; KL, $P=6.53\times10^{-1}$, 3.14×10^{-1}, 5.53×10^{-5}, 1.17×10^{-5}, 2.71×10^{-8} and 5.75×10^{-5}, respectively; KW, $P=2.50\times10^{-1}$, 6.41×10^{-1}, 9.32×10^{-1}, 5.72×10^{-1}, 2.85×10^{-2} and 2.07×10^{-4}, respectively; ED, $P=1.19\times10^{-2}$, 1.60×10^{-1}, 7.03×10^{-3}, 1.73×10^{-2}, 2.97×10^{-2} and 2.99×10^{-1}, respectively; KRN, $P=2.06\times10^{-2}$, 9.68×10^{-1}, 3.30×10^{-2}, 1.78×10^{-1}, 2.00×10^{-1} and 2.95×10^{-1}, respectively; CW, $P=1.93\times10^{-1}$, 1.38×10^{-1}, 3.98×10^{-1}, 4.32×10^{-2}, 2.71×10^{-3} and 8.35×10^{-1}, respectively. **d**, Phenotypic changes in the FHGs and MHGs and in the two female groups (PA and US_SS) and four male groups (PB, SPT, CN_NSS and US_NSS) from Era I to Era II for the three traits of Group 4. The P values of FHGs[Era I] versus FHGs[Era II], MHGs[Era I] versus MHGs[Era II], FHGs[Era I] versus MHGs[Era I] and FHGs[Era II] versus MHGs[Era II] using two-sided Wilcoxon test are presented in Supplementary Table 5. The P values of Era I versus Era II in six groups using two-sided Wilcoxon test (left to right, as plotted) are: PH, $P=1.97\times10^{-1}$, 3.34×10^{-1}, 8.63×10^{-2}, 1.52×10^{-1}, 1.57×10^{-1} and 6.20×10^{-1}, respectively; TL, $P=2.87\times10^{-1}$, 3.53×10^{-1}, 3.45×10^{-1}, 3.78×10^{-3}, 8.49×10^{-1} and 5.09×10^{-2}, respectively; EL, $P=9.79\times10^{-1}$, 7.53×10^{-1}, 2.03×10^{-1}, 7.36×10^{-1}, 5.80×10^{-1} and 9.33×10^{-5}, respectively. For boxplots in **a-d**, the central lines show the median, the box limits indicate the 25th and 75th percentiles, whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles. Exact sample size (n) in FHGs[Era I], FHGs[Era II], MHGs[Era I] and MHGs[Era II] (left to right, as plotted): n = 72, 253, 330 and 267. Exact sample size (n) in PA[Era I], PA[Era II], US_SS[Era I], US_SS[Era II], PB[Era I], PB[Era II], SPT[Era I], SPT[Era II], CN_NSS[Era I], CN_NSS[Era II], US_NSS[Era I] and US_NSS[Era II] (left to right, as plotted): n = 46, 161, 26, 92, 32, 42, 46, 78, 204, 86, 48 and 61. Significant differences are indicated: ** $P < 0.01$ and * $P < 0.05$.
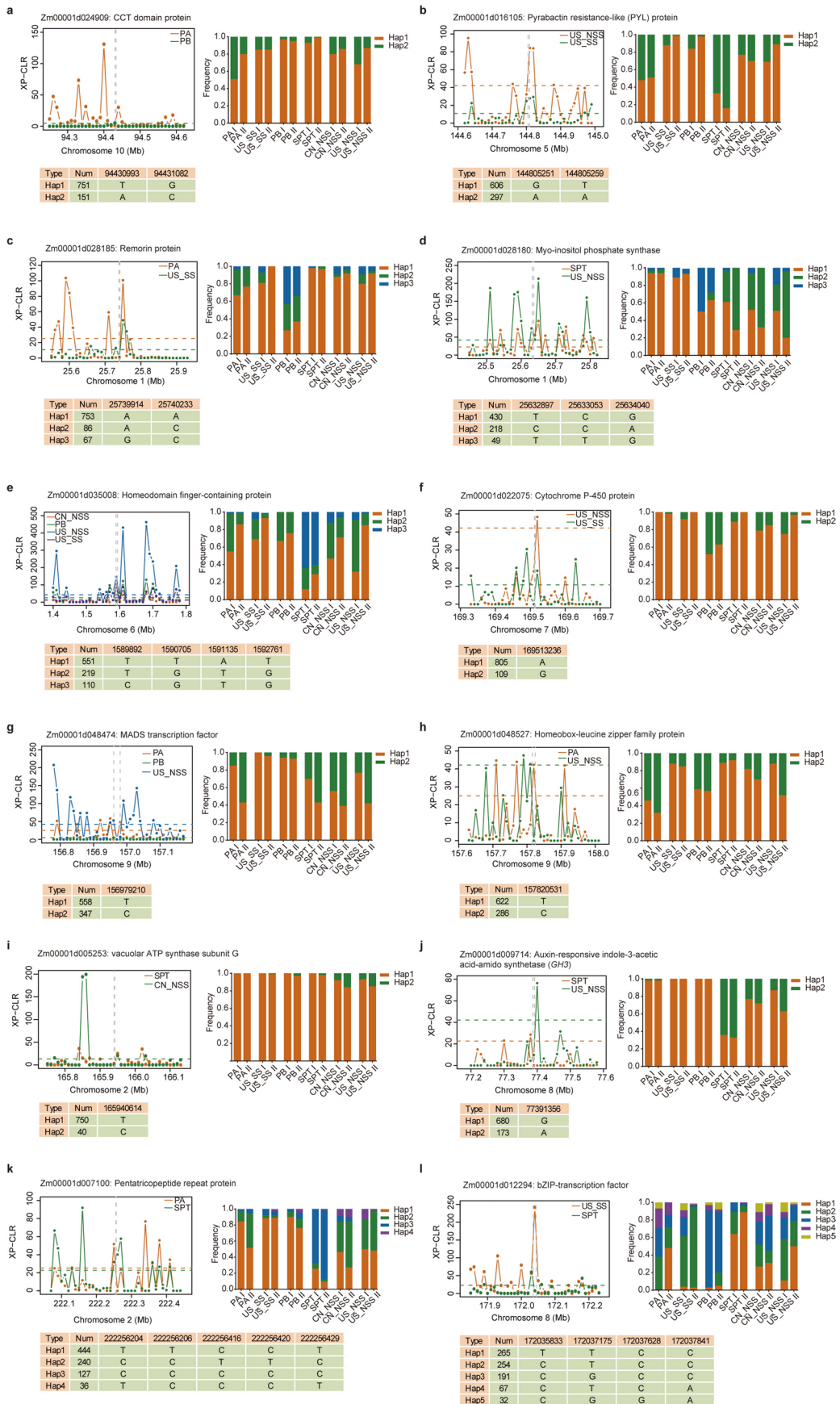
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Accumulation of favorable alleles in the female and male heterotic groups during modern hybrid maize breeding. a, Profile of favorable allele frequencies changes at GWAS associated SNPs in the FHGs and MHGs. Orange indicates an increase, while green indicates a decrease in the frequency of a favorable allele from Era I to Era II during modern breeding. Each row represents an associated SNP. Blue and gray colors (in the first column) mark rows representing the associated SNPs at the threshold of $P < 10^{-6}$ and 10^{-5}, respectively. **b**, The number of favorable alleles accumulated in the FHGs and MHGs across different breeding eras. The P values of FHG$^{Era\,I}$ versus FHG$^{Era\,II}$, MHG$^{Era\,I}$ versus MHG$^{Era\,II}$, FHG$^{Era\,I}$ versus MHG$^{Era\,I}$ and FHG$^{Era\,II}$ versus MHG$^{Era\,II}$ using two-sided Wilcoxon test are: DA, $P = 3.70 \times 10^{-1}$, 3.71×10^{-2}, 2.10×10^{-1} and 2.04×10^{-4}, respectively; DS, $P = 5.60 \times 10^{-2}$, 3.62×10^{-4}, 3.36×10^{-1} and 7.53×10^{-1}, respectively; ASI, $P = 1.73 \times 10^{-1}$, 2.54×10^{-15}, 1.42×10^{-13} and 9.62×10^{-7}, respectively; EH, $P = 4.12 \times 10^{-1}$, 3.52×10^{-1}, 7.55×10^{-1} and 2.63×10^{-2}, respectively; EP, $P = 5.15 \times 10^{-1}$, 1.77×10^{-1}, 8.08×10^{-1} and 1.50×10^{-1}, respectively; TBN, $P = 8.11 \times 10^{-7}$, 2.68×10^{-8}, 2.34×10^{-19} and 4.83×10^{-41}, respectively; KWPE, $P = 5.30 \times 10^{-1}$, 9.80×10^{-18}, 1.70×10^{-15} and 2.63×10^{-10}, respectively; EW, $P = 6.43 \times 10^{-1}$, 3.22×10^{-15}, 2.61×10^{-15} and 9.41×10^{-10}, respectively; KNPR, $P = 2.67 \times 10^{-1}$, 3.74×10^{-15}, 1.15×10^{-17} and 7.94×10^{-17}, respectively; HKW, $P = 6.51 \times 10^{-2}$, 2.50×10^{-13}, 9.47×10^{-5} and 3.28×10^{-3}, respectively; HKV, $P = 2.22 \times 10^{-1}$, 1.44×10^{-13}, 4.83×10^{-4} and 7.16×10^{-2}, respectively; KL, $P = 2.43 \times 10^{-1}$, 8.35×10^{-23}, 4.71×10^{-17} and 5.82×10^{-3}, respectively; KW, $P = 8.62 \times 10^{-1}$, 3.82×10^{-7}, 2.95×10^{-1} and 1.65×10^{-4}, respectively; KRN, $P = 2.81 \times 10^{-2}$, 2.16×10^{-1}, 1.19×10^{-8} and 4.32×10^{-4}, respectively; CW, $P = 9.11 \times 10^{-1}$, 7.04×10^{-3}, 1.09×10^{-5} and 1.20×10^{-5}, respectively. For boxplots, the central lines show the median, the box limits indicate the 25th and 75th percentiles, whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles. Significant differences are indicated: *** $P < 0.001$, ** $P < 0.01$ and * $P < 0.05$. Exact sample size (n) are 72, 253, 330 and 267 for FHGs$^{Era\,I}$, FHGs$^{Era\,II}$, MHGs$^{Era\,I}$ and MHGs$^{Era\,II}$, respectively. **c**, Correlation between the number of accumulated favorable alleles and improved trait in the FHGs and MHGs. The six traits of Group 1 include DA (days to anthesis), DS (days to silking), ASI (anthesis to silking interval), EH (ear height), EP (relative ear height), TBN (tassel branch number); the seven traits of Group 2 include KWPE (kernel weight per ear), EW (ear weight), KNPR (kernel number per row), HKW (hundred kernel weight), HKV (hundred kernel volume), KL (kernel length) and KW (kernel width); and the two traits of Group 3 include KRN (kernel row number) and CW (cob weight). The Pearson correlation coefficient (r) and P value are presented.
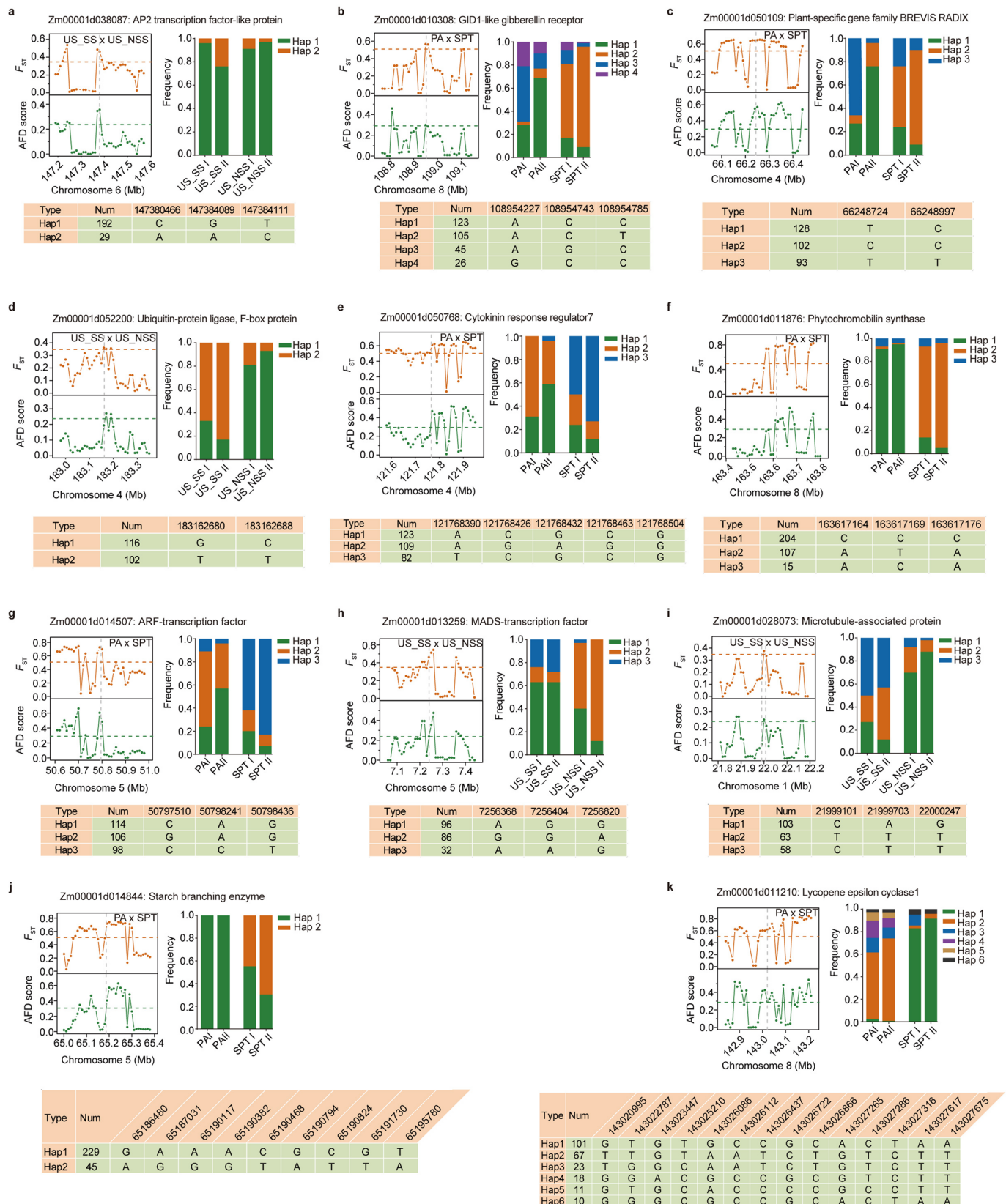
Extended Data Fig. 4 | Distribution and sharing of selective sweeps and profile of allele frequency change for the selected genes in the FHGs and MHGs. a, The distribution ratio of selective sweeps located in genic and intergenic regions in the six heterotic groups. **b**, Distribution of the distances from intergenic selective sweeps to the nearest gene. **c**, Enrichment of shared genes (below the diagonal) and GO terms (above the diagonal) in the female heterotic groups (FHGs) and/or male heterotic groups (MHGs). Squares are marked with an asterisk (P < 0.05) and two asterisks (P < 0.01) if the observed number shared was significantly higher than the background number under permutation test conditions. **d**, Profile of allele frequency change of the 589 selected genes exhibiting co-directional change in allele frequencies between the FHGs and MHGs from Era I to Era II. **e**, Profile of allele frequency change of the 28 selected genes exhibiting anti-directional changes between the FHGs and MHGs from Era I to Era II. **f**, Profile of allele frequency change of the 400 selected genes exhibiting convergent increase or reduction in allele frequencies in the MHGs, but not in the FHGs, from Era I to Era II. For **d-f**, Orange indicates an increase of the reference allele frequency of nonsynonymous SNPs, whereas green indicates an increase of the alternative allele frequency from Era I to Era II. Each row represents a nonsynonymous SNP of the selected genes.
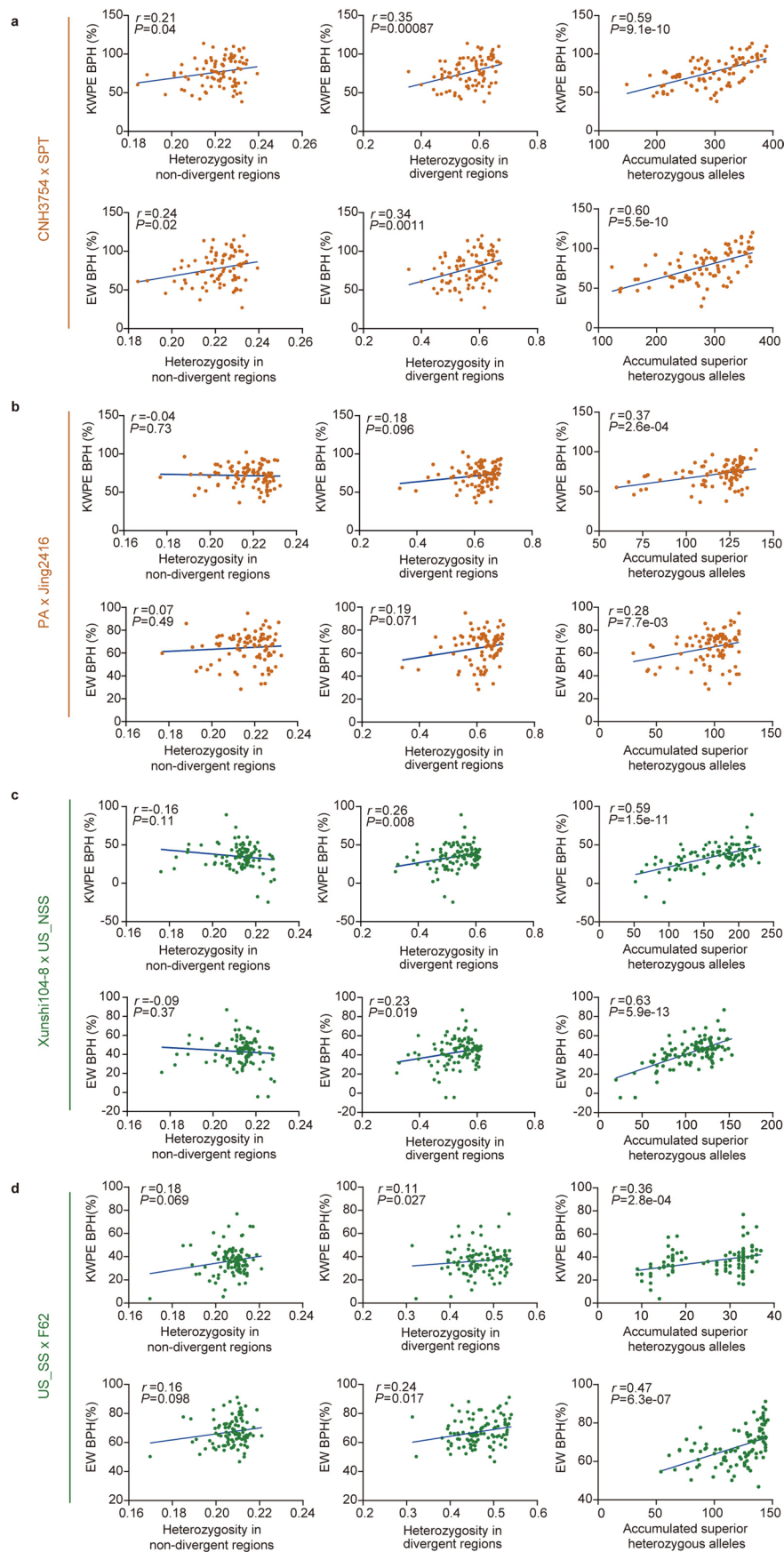
a Zm00001d024909: CCT domain protein

b Zm00001d016105: Pyrabactin resistance-like (PYL) protein

c Zm00001d028185: Remorin protein

d Zm00001d028180: Myo-inositol phosphate synthase

e Zm00001d035008: Homeodomain finger-containing protein

f Zm00001d022075: Cytochrome P-450 protein

g Zm00001d048474: MADS transcription factor

h Zm00001d048527: Homeobox-leucine zipper family protein

i Zm00001d005253: vacuolar ATP synthase subunit G

j Zm00001d009714: Auxin-responsive indole-3-acetic acid-amido synthetase (GH3)

k Zm00001d007100: Pentatricopeptide repeat protein

l Zm00001d012294: bZIP-transcription factor

Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Representative selected genes related to plant growth and development and abiotic stress responses in at least two heterotic groups. a-l, Genes functionally characterized in maize or their homologous genes have been functionally characterized in rice. Each gene includes three plots: XP-CLR plot, physical position of the gene (indicated by vertical gray dotted line), top 5% score of XP-CLR in individual heterotic group (shown by horizontal dotted line). Haplotype table plot, the type and number of haplotypes formed by nonsynonymous SNPs of gene were counted. The haplotypes with at least 20 inbred lines were listed in the table. Haplotype frequency bar plot, different haplotype frequency change from Era I to Era II was shown within each of the six heterotic groups.
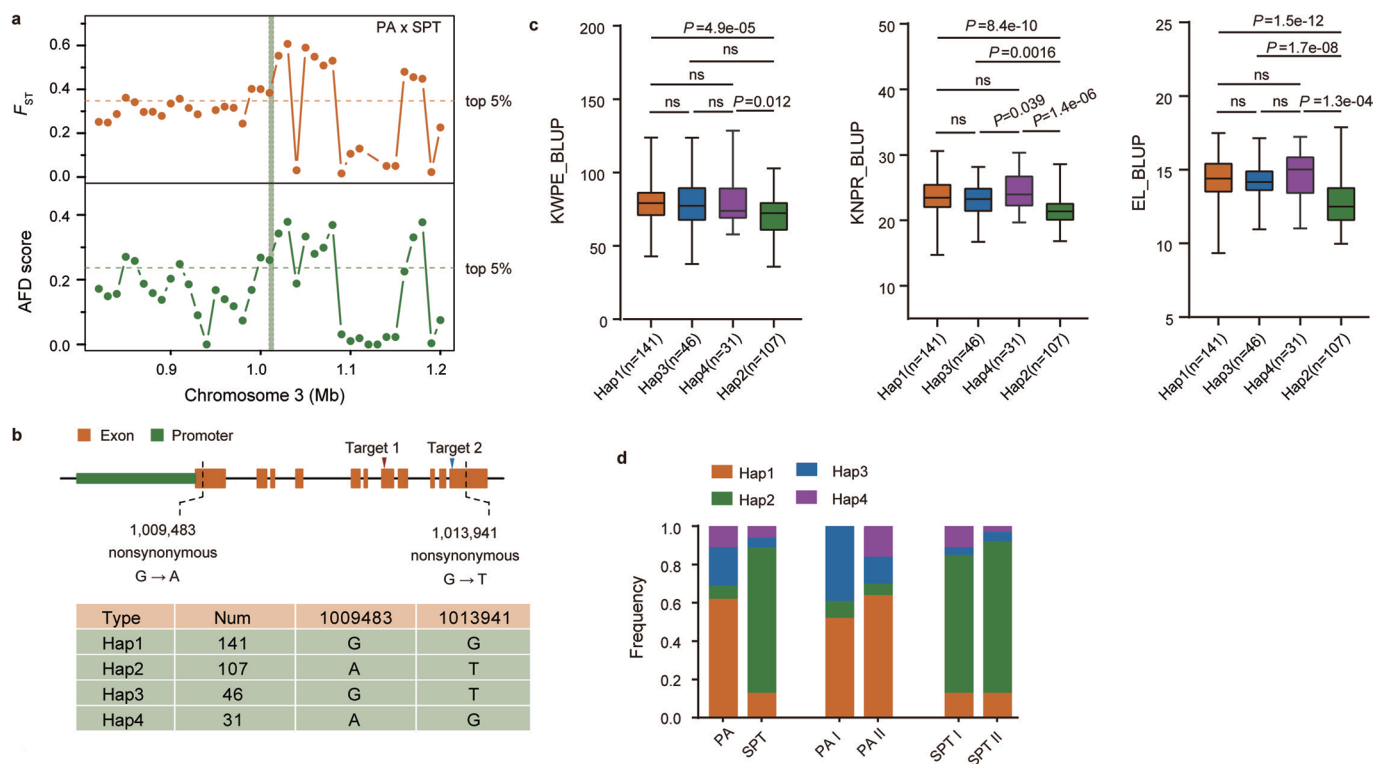
Extended Data Fig. 6 | Representative differentiated genes related to abiotic stress responses and plant growth and development. a–k, Genes functionally characterized in maize or their homologous genes have been functionally characterized in rice. Each gene includes three plots: F_{ST} and allele frequency difference (AFD) plot, physical position of the gene (indicated by vertical gray dotted line), top 5% score of F_{ST} and AFD between the FHGs and MHGs were shown by horizontal orange dotted line and green dotted line, respectively. Haplotype table plot, the type and number of haplotypes formed by nonsynonymous SNPs of gene were counted. The haplotypes with at least 10 inbred lines were listed in the table. Haplotype frequency bar plot, different haplotype frequency change from Era I to Era II was shown in the female and male heterotic groups.

Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | The relationships between heterozygosity levels and heterosis and between accumulated superior heterozygous alleles and heterosis for yield-related traits in four testcross populations. a-d, Each plot group includes six plots for two traits (kernel weight per ear (KWPE) and ear weight (EW)): Left plot, correlation between heterozygosity levels of whole-genome nonsynonymous SNPs in non-divergent regions and better parent heterosis (BPH) of KWPE and EW in the testcross population. Middle plot, correlation between heterozygosity levels of nonsynonymous SNPs located in genes continuously selected in divergent regions and BPH of KWPE and EW in the testcross population. A total of 2,563 and 2,136 nonsynonymous SNPs contained in 478 and 375 genes continuously selected in PA × SPT and US_SS × US_NSS, were used to calculate the heterozygosity levels for the testcross populations from PA × SPT and SS × NSS heterotic patterns, respectively. Right plot, correlation between the number of accumulated superior heterozygous alleles and BPH of KWPE and EW in the testcross population. Four testcross populations include 88 hybrids derived from CNH3754 (a PA inbred) × 88 SPT inbreds, 91 hybrids derived from 91 PA inbreds × Jing2416 (a SPT inbred), 106 hybrids derived from Xunshi104-8 (a SS inbred) × 106 US_NSS inbreds, and 101 hybrids derived from 101 US_SS inbreds × F62 (a NSS inbred). The Pearson correlation coefficient (*r*) and *P* value are presented.

Extended Data Fig. 8 | Identification of ZmKOB1 as a differentiated gene between the PA and SPT heterotic groups. a, F_{ST} (above the axis) and allele frequency difference (AFD) of *ZmKOB1*. The candidate gene was visualized by green shadow. **b**, Gene structure and haplotype analyses of *ZmKOB1*. **c**, Box plots for kernel weight per ear (KWPE), kernel number per row (KNPR) and ear length (EL) for the four haplotypes. Center line, medium; box limits, upper and lower quartiles; whiskers, 1.5 × interquartile range. n indicates the number of inbred lines for each haplotype. The significance of difference was analyzed using two-sided Wilcoxon test. **d**, Haplotype frequency of *ZmKOB1* in different breeding eras in the PA and SPT heterotic groups.

nature portfolio

Corresponding author(s): Haiyang Wang

Last updated by author(s): Jun 5, 2022

Reporting Summary

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

Statistics

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

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We used open source software and codes for data collection, including BWA (v0.7.17-r1188), SAMtools (v1.3), Sambamba (v0.6.8), GATK(v4.1.2.0), ANNOVAR, fastStructure (v1.0), TreeBest (v1.9.2), PopLDdecay, XP-CLR, VCFtools (v0.1.15), Beagle (4.1), R (v3.6.1), EMMAX, agriGO (v2.0) and ClusterProfiler (v3.10.0). The custom codes used in this study are available at: https://github.com/jasongit0311/maize_for_Li.
Data analysis	All softwares used in the present study are publicly available and the corresponding versions are described in detail in the Online Methods.

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

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

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

The raw sequencing data were deposited in the Genome Sequence Archive (https://bigd.big.ac.cn/gsa) under Accession Code PRJCA009749. The phenotype dataset reported here is available from the corresponding authors upon request. Source data are provided with this paper.

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	A total of 1,604 maize inbred lines, which comprise of 1,227 inbreds from China, 344 inbreds from the United States, and 33 inbreds from other countries, were collected for this study. No sample size calculation was performed. These samples were very sufficient to study the heterotic groups improvement during modern breeding, and at least 26 samples (average about 76) were used for each breeding era in individual heterotic group.
Data exclusions	During the XP-CLR analysis, to eliminate the interference of small samples (less than 20 inbreds) within each breeding era, we excluded those inbreds from two breeding eras in Iodent heterotic group in the selective sweep analysis. This was clearly described in the manuscript.
Replication	The 21 agronomic traits for 1,604 inbred lines were repeatedly measured across ten environments. Ten sampling replicates were used for expression analysis of ZmEMF1L1 and ZmKW10, with each replicate consisting of leaf or kernel tissues from 3 independent plants.
Randomization	A randomized complete block design was used in all trials for phenotype collection.
Blinding	the investigators were blinded to group allocation during data analysis.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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