nature genetics

LETTERS

https://doi.org/10.1038/s41588-020-0616-3

Check for updates

Genome-wide selection and genetic improvement during modern maize breeding

Baobao Wang[1,13], Zechuan Lin[2,13], Xin Li[1,13], Yongping Zhao[1], Binbin Zhao[1], Guangxia Wu[1], Xiaojing Ma[1], Hai Wang[1], Yurong Xie[1], Quanquan Li[1,3], Guangshu Song[4], Dexin Kong[5], Zhigang Zheng[5], Hongbin Wei[5], Rongxin Shen[5], Hong Wu[5], Cuixia Chen[3], Zhaodong Meng[6], Tianyu Wang[7], Yu Li[7], Xinhai Li[7], Yanhui Chen[8], Jinsheng Lai[9], Matthew B. Hufford[10], Jeffrey Ross-Ibarra[11], Hang He[2] ✉ and Haiyang Wang[5,12] ✉

Since the development of single-hybrid maize breeding programs in the first half of the twentieth century[1], maize yields have increased over sevenfold, and much of that increase can be attributed to tolerance of increased planting density[2–4]. To explore the genomic basis underlying the dramatic yield increase in maize, we conducted a comprehensive analysis of the genomic and phenotypic changes associated with modern maize breeding through chronological sampling of 350 elite inbred lines representing multiple eras of germplasm from both China and the United States. We document several convergent phenotypic changes in both countries. Using genome-wide association and selection scan methods, we identify 160 loci underlying adaptive agronomic phenotypes and more than 1,800 genomic regions representing the targets of selection during modern breeding. This work demonstrates the use of the breeding-era approach for identifying breeding signatures and lays the foundation for future genomics-enabled maize breeding.

Maize (*Zea mays* ssp. *mays* L.) is a major staple crop worldwide, accounting for 37.2% of total worldwide cereal production[5]. Previous studies have investigated the genome-wide changes that occurred during maize domestication[6–8] and expansion to novel environments[9–12] but analysis of genetic improvement during modern breeding has so far been limited in scope[13–16]. To investigate the genetic impacts of selection during modern maize breeding and identify the key genes contributing to adaptation to increased planting density, we collected 350 elite maize inbred lines, including 163 inbred lines from the United States and 187 inbred lines from China. The US inbred lines comprise 74 public (hereafter Public-US) and 89 elite commercial lines with expired Plant Variety Protection Act Certificates mostly released after 2003 (hereafter Ex-PVP). The Chinese inbred lines are divided into three groups on the basis of their date of release and use in hybrid breeding: 30 early-stage inbred lines (released during 1960–1979, hereafter CN1960&70s), 95 middle-stage inbred lines (released during 1980–1999, hereafter CN1980&90s) and 53 recently released elite inbred lines (released after 2000, hereafter CN2000&10s; Fig. 1a and Supplementary Table 1).

We first phenotyped 15 key agronomic traits for 2 consecutive years across four locations (Supplementary Table 2). In agreement with earlier reports[16,17], we observed convergent phenotypic selection in the United States and Chinese inbred lines for three agronomic traits during the modern breeding process, that is reductions in upper leaf angle (LAU), tassel branch number (TBN) and anthesis-silking interval (ASI; Fig. 1). These traits were thought to be important for adaptation to high planting density[4]. Significant changes ($P < 0.05$) in flowering time (days to silking (DTS) and days to anthesis (DTA)) and the relative height of the ear (EP) were also observed, while traits such as plant height (PH) showed no noticeable change (Fig. 1 and Extended Data Fig. 1). Similar results were also observed within individual heterotic pools (Supplementary Fig. 1).

To characterize the genetic basis of these phenotypic changes, we sequenced the 350 inbred lines to an average depth of 13.4× (10.0–28.9×; Supplementary Table 1) and identified >25,000,000 high-quality single-nucleotide polymorphisms (SNPs) and >4,000,000 small (<10 base pairs (bp)) indels following a strict filtering pipeline (Supplementary Tables 3 and 4). In agreement with previous genotyping results[14], we found decreased nucleotide diversity and increased linkage disequilibrium (LD) in the US Ex-PVP lines when compared to the Public-US lines (Fig. 2). Similar patterns were observed between CN2000&10s and CN1960&70s. However, no significant difference was observed between the CN1960&70s and CN1980&90s inbred lines (Fig. 2 and Supplementary Fig. 1). Population structure analysis revealed that these inbred lines could be grouped into four main groups corresponding to the Stiff Stalk Synthetic (SS), Nonstiff Stalk (NSS), Iodent (IDT) and the China-specific group Huangzaosi (HZS; Fig. 2 and Supplementary

[1]Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing, China. [2]College of Life Sciences, State Key Laboratory of Protein and Plant Gene Research, Peking-Tsinghua Center for Life Sciences, School of Advanced Agricultural Sciences, Peking University, Beijing, China. [3]State Key Laboratory of Crop Biology, Shandong Agricultural University, Taian, China. [4]Maize Research Institute, Jilin Academy of Agricultural Sciences, Gongzhuling, China. [5]College of Life Sciences, State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, South China Agricultural University, Guangzhou, China. [6]Maize Research Institute, Shandong Academy of Agricultural Sciences, Jinan, China. [7]Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China. [8]Synergetic Innovation Centre of Henan Grain Crops and National Key Laboratory of Wheat and Maize Crop Science, Henan Agricultural University, Zhengzhou, China. [9]State Key Laboratory of Agrobiotechnology and National Maize Improvement Center, Department of Plant Genetics and Breeding, China Agricultural University, Beijing, China. [10]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA, USA. [11]Department of Evolution and Ecology, Center for Population Biology, and Genome Center, University of California, Davis, CA, USA. [12]Guangdong Laboratory for Lingnan Modern Agriculture, Guangzhou, China. [13]These authors contributed equally: Baobao Wang, Zechuan Lin, Xin Li. ✉e-mail: hehang@pku.edu.cn; whyang@scau.edu.cn

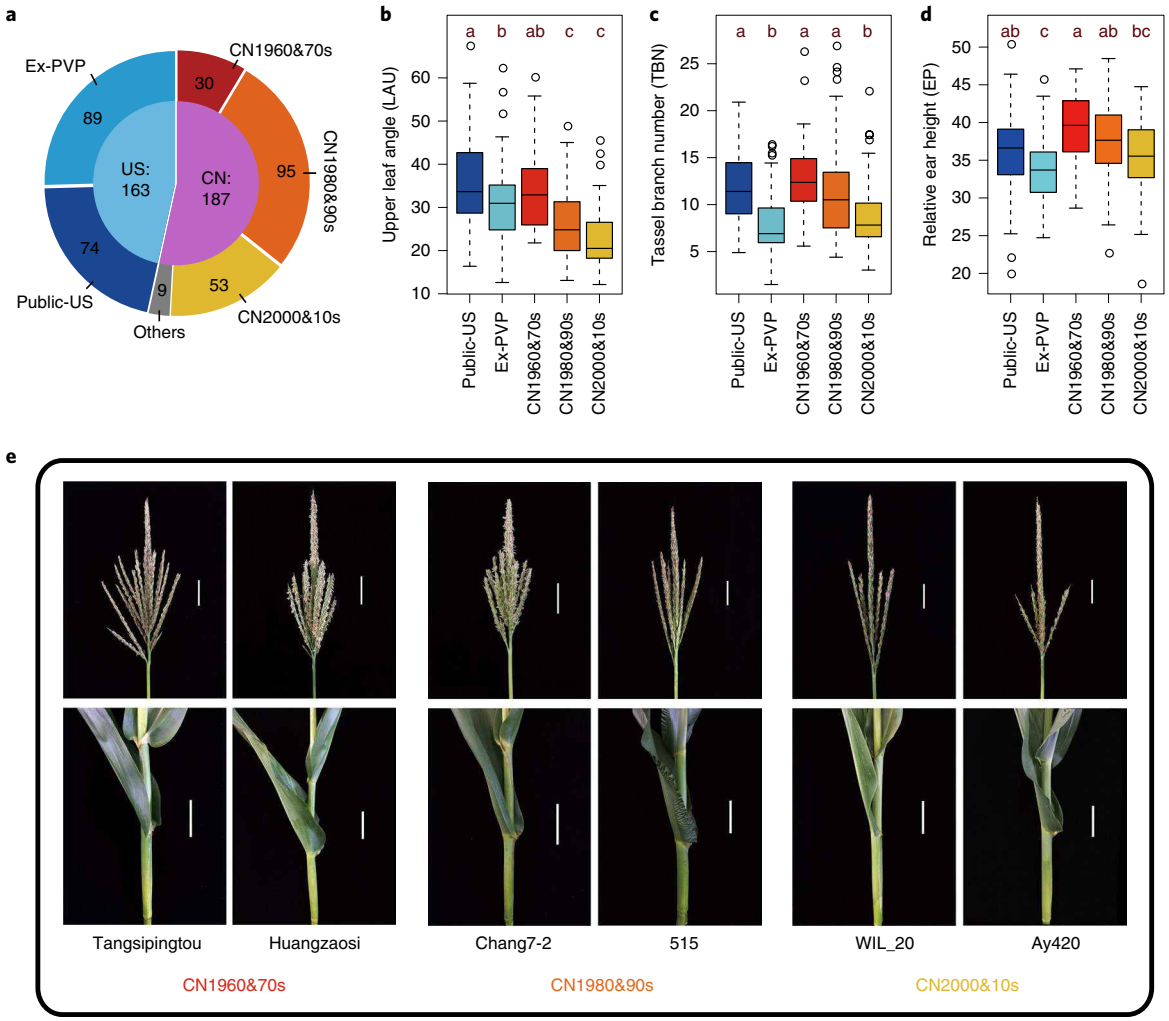


Fig. 1 | Morphological trait improvement during modern maize breeding in the United States and China. a, Distribution of 350 elite inbred lines of different breeding eras. **b–d**, Phenotypic distributions of LAU (**b**), TBN (**c**) and EP (**d**) among different breeding eras in the United States and China. For each box, the upper and lower boundaries represent the 25th and 75th percentile, respectively. The middle horizontal lines represent the median. The whiskers represent 1.5× the interquartile range. The dots beyond the whiskers represent outliers. Different letters above the boxes indicate significant differences ($P < 0.05$, Bonferroni correction) in a pairwise comparison. **e**, Representative images of tassel and leaf angle of typical inbred lines in the HZS group during the Chinese maize breeding history. Scale bars, 5 cm.

Table 5), in close agreement with previous reports and the known introduction history of maize germplasm used in the United States and China[14,18].

We performed genome-wide association study (GWAS) for each of the 15 key agronomic traits and identified a total of 233 significant loci ($P < 1 \times 10^{-6}$, false discovery rate (FDR) < 0.05; Supplementary Table 6). We successfully identified *Pericarp color1* (*P1*)[19], *Yellow endosperm1* (*Y1*)[20], *White Cap1* (*WC1*)[21] and *Vegetative to Generative Transition1* (*VGT1*)[22] (Extended Data Fig. 2), validating the effectiveness of our approach. A total 128 of our loci were located within ~1 megabase (Mb) of previously reported quantitative trait nucleotides (QTNs) for 14 traits (Supplementary Table 7). We also identified some additional highly promising associations. For example, both *ZmNAC16* (GRMZM2G166721) and *ZmSBP18* (GRMZM2G371033) were associated with leaf angle, in agreement with studies of their homologs in other grasses[23–25]. For both genes, qRT–PCR analysis revealed differential expression in collars of expanded leaves for inbred lines with contrasting haplotypes (Fig. 3a–h). *ZmRVE1* (GRMZM2G181030), encoding a MYB family transcription factor homologous to *Arabidopsis REVEILLE 1* (*RVE1*)[26,27], is associated with DTA. Consistent with the documented role of cytokinin in regulating TBN[28,29], *ZmCRF4* (GRMZM2G142179, encoding a protein homologous to *Arabidopsis* CRF4) and *ZmARR2* (GRMZM2G126834, encoding a protein homologous to *Arabidopsis* ARR2) are associated with TBN. For each of these loci, putative causal polymorphisms were identified with changes in frequency of the favorable allele consistent with selection during modern breeding (Fig. 3).

To test for evidence of selection on agronomic phenotypes, we asked whether the favorable allele (alleles associated with reduced ear height, more erect leaves, reduced TBN and accelerated flowering) at each associated SNP increased in frequency over time during the process of breeding. We found evidence of convergent increases in allele frequency at putatively favorable alleles for 41.7% of loci for EP, 66.2% for lower leaf angle (LAL), 64.1% for LAU and 49.5% for TBN in both the United States and China (Supplementary Table 8 and Extended Data Fig. 3) and most loci for EP, LAL, LAU, TBN, DTS and DTA showed evidence of selection in either China or the United States. Although the magnitude of increase was in most cases small, with an average increase of ~0.110 (ranging from 0.001 to 0.590 in China and from 0.001 to 0.410 in the United States), this was significantly greater than expected by chance (permutation

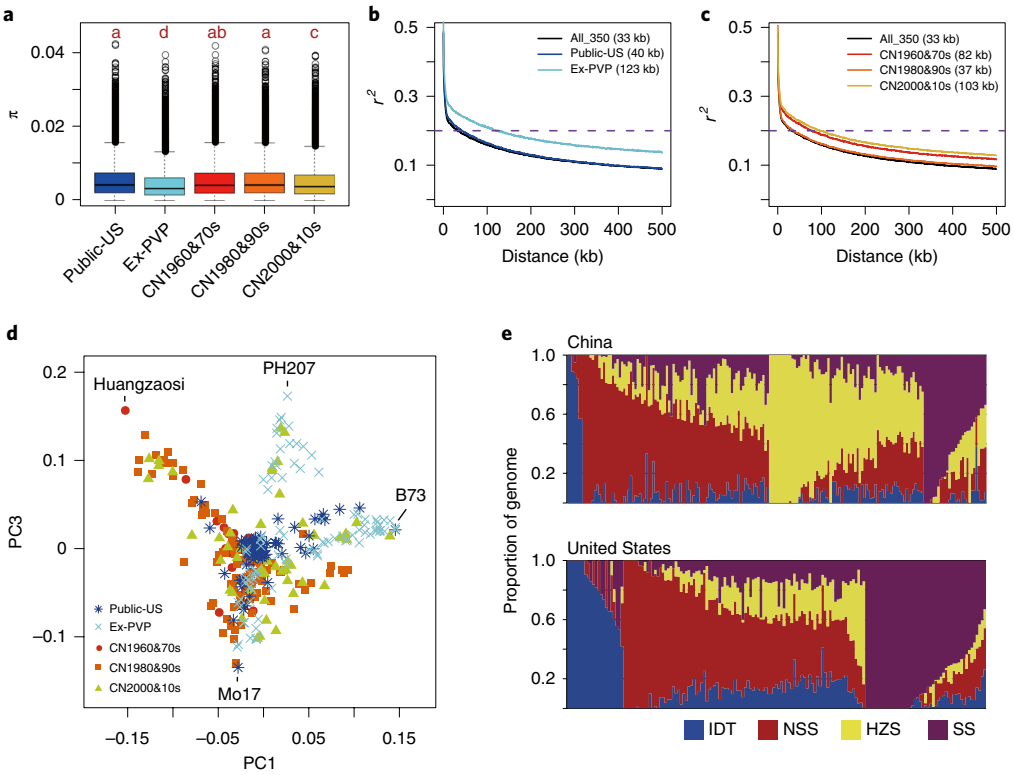


Fig. 2 | Nucleotide diversity, LD decay and population structure of 350 maize inbred lines. a, Nucleotide diversity of inbred lines from different maize breeding eras in the United States and China. **b**,**c**, Genome-wide averaged distance of LD decayed to $r^2 = 0.2$ for different US inbred lines (**b**) and Chinese inbred lines (**c**). **d**, PCA plots of the first and third PCs. Inbred lines from different eras are represented in different colors and shapes as shown. **e**, Population structure of US and Chinese maize inbred lines at prior number of ancestor populations ($K = 4$). The inbred lines are classified into four canonical groups corresponding to: IDT, represented by PH207; NSS, represented by Mo17; HZS, represented by HZS; and SS, represented by B73.

$P < 0.03$; Supplementary Table 9) and is consistent with models of selection on polygenic traits[30]. The convergent selection of these traits was also supported by analyses of polygenic scores, gene flow and the relative frequencies of the most predictive SNPs in our materials (Supplementary Note and Supplementary Table 10).

Many loci targeted by selection during modern breeding may not directly contribute to an obvious phenotype[31]. We thus used the cross-population composite likelihood ratio approach (XP-CLR)[32] to detect putative selected regions of different maize breeding eras: CN1980&90s versus CN1960&70s, CN2000&10s versus CN1980&90s and Ex-PVP versus Public-US. To study long-term selection during breeding in China, we also compared CN2000&10s versus CN1960&70s. After removal of candidate regions potentially driven by population structure, a total of 1,888 selected regions were detected in at least one of the comparisons (Fig. 4a and Supplementary Table 11). These regions show a greater reduction in nucleotide diversity and greater differentiation than the rest of the genome (Supplementary Table 12). Combined across eras and populations, selected regions encompassed 5,356 genes and comprised 13.64% of the maize genome. Nonetheless, individual selected regions were fairly small, with mean sizes ranging from 121 to 183 kilobases (kb) (Supplementary Table 12). These sweeps were smaller than previously observed for domestication (average size 322 kb) but comparable to sweeps for improvement (average size 176 kb)[8] and tropical–temperate adaptation (average size 150.9 kb)[12]. Notably, only limited overlap (~9.60% on average) was found between our identified selective sweeps and previously reported domestication and temperate adaptation sweeps[8,12] (Supplementary Table 12), indicating that the genomic regions selected during modern breeding are distinct from selection during earlier periods of maize evolution.

Gene ontology analysis of the 5,356 genes encompassed in the selective sweeps revealed enrichment of genes in responses to biotic and abiotic stress, response to light, biosynthesis or signaling processes of auxin and other phytohormones (Extended Data Figs. 4–7, Supplementary Fig. 2 and Supplementary Tables 13–15). We further identified a subset of 2,009 genes containing nonsynonymous variants that showed significant change in allele frequency across the breeding eras (Fisher's exact test, $P < 0.05$; Supplementary Table 16). Since nonsynonymous mutations may cause adaptive alteration in the encoded proteins and changes in allele frequency may reflect selection during breeding, we viewed these genes as particularly interesting candidates of selection.

In agreement with our results showing convergent phenotypic change and selection of associated alleles, we found evidence of genome-wide parallel selection between China and the United States, with 304 sweeps (encompassing 724 genes) shared between the United States and at least one of the comparisons in China ($P \cong 0$, one-tail t-test; Fig. 4b,c). Our data also provide evidence of sustained selection at individual genomic regions, with, for example, 98 sweeps (including 281 genes) targeted during both early and late periods of Chinese breeding ($P \cong 0$, one-tail t-test). Nonetheless, many distinct candidate regions were identified in individual scans, potentially due to environmental differences, local breeding preferences, or shifting selection pressure over time.

The intersection of genome scan and association mapping results points to loci of particular interest for modern breeding. In total, 41 of our GWAS loci overlapped with regions showing evidence of selection. Notably, 61% (25/41) of them were loci for plant architecture traits important for modern breeding (two for EH, five for EP, three for LAU, five for LAL and ten for TBN; Supplementary Table 17). In the overlapping regions, we observed two associated loci

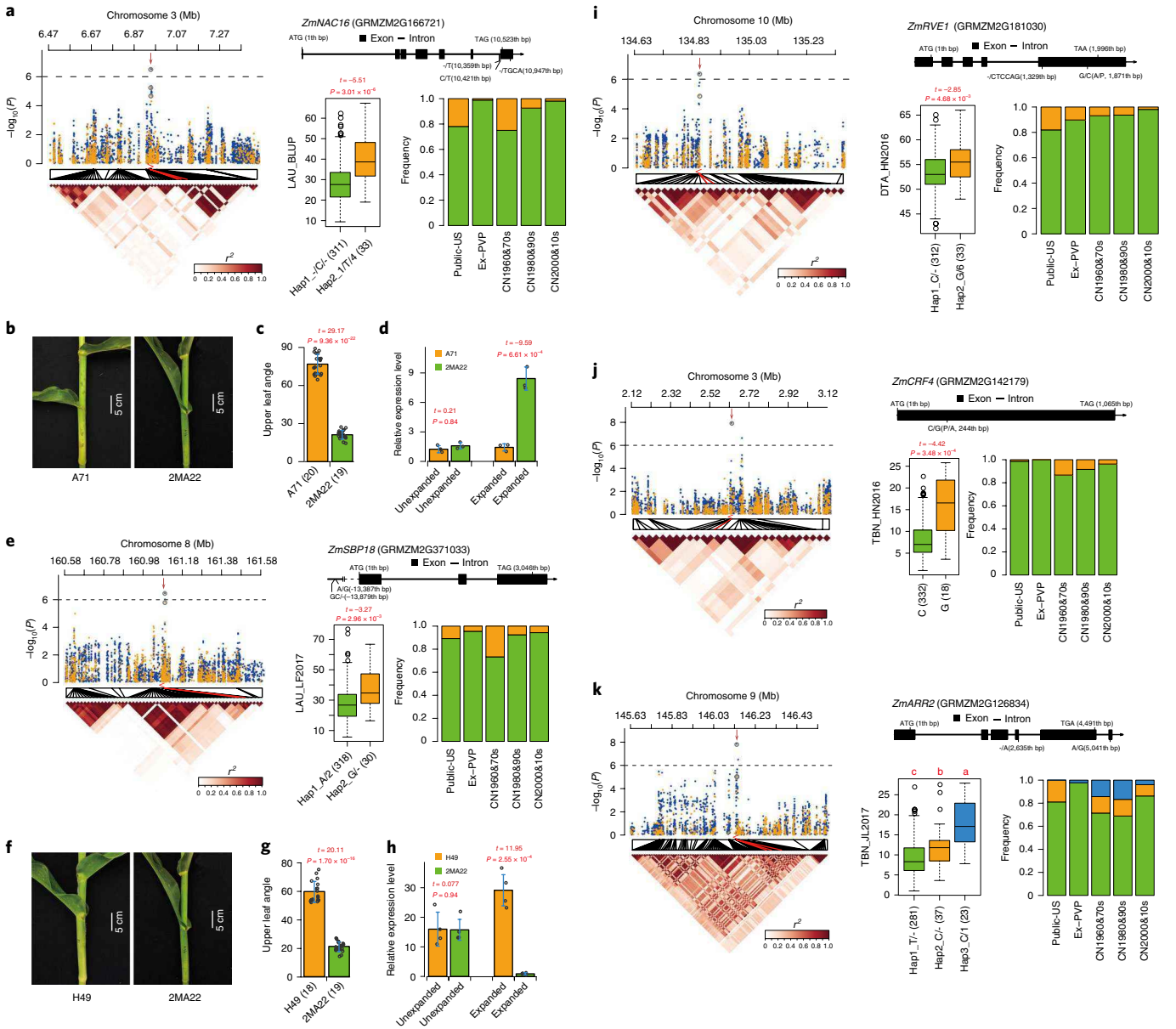


Fig. 3 | GWAS identification of candidate genes for variation of LA, DTA and TBN. a, GWAS identification of *ZmNAC16* as a candidate gene for LA variation. The group of plots includes a Manhattan plot (upper left) and LD heat map (lower left), candidate gene structure and putative causal polymorphisms (upper right), phenotype of different haplotypes (box plot) and haplotype frequency changes during breeding (bar plot), for the GWAS signal of LAU_3_6945310. SNP- and indel-based association analysis results are shown as blue and orange dots in the Manhattan plot, respectively. Peak markers and putative causal polymorphisms are circled and their positions in the LD heat map are indicated by red lines. The candidate gene position in Manhattan plot is showed as red arrows. The *t* value and *P* value for two-tailed *t*-test are shown above the box plot. Sample numbers are shown in brackets on *x* axis. **b**,**c**, Images (**b**) and statistics (**c**) of leaf angle of the two inbred lines A71 and 2MA22. **d**, qRT-PCR analysis of *ZmNAC16* expression in the leaf collars of A71 and 2MA22: error bars ± s.d.; $n = 3$. **e**, GWAS identification of *ZmSBP18* as a candidate gene for LA variation. **f**,**g**, Images (**f**) and statistics (**g**) of leaf angle of H49 and 2MA22. **h**, qRT-PCR analysis of *ZmSBP18* expression in the leaf collars of H49 and 2MA22; $n = 4$. **i**, GWAS identification of *ZmRVE1* as a candidate gene for DTA variation. **j**,**k**, GWAS identification of *ZmCRF4* (**j**) and *ZmARR2* (**k**) as candidate genes for TBN variation.

for EP (Extended Data Fig. 8). One peak SNP is located in the genic region of GRMZM2G398996, which encodes a protein homologous to the gibberellin receptor GID1-like 2, which is known to regulate plant height and architecture[33]. Another locus (EP_1_25433596) shows evidence of selection in the late breeding process in China and is located in a region with a large LD block (>1 Mb).

To validate the approaches we used to identify loci important for modern breeding, we identified a high-confidence candidate gene from each approach and assayed the phenotypic effects of CRISPR–Cas9 knockout lines. Phytochromes are known to play an important role in regulating maize architecture and flowering time, and are likely targets of selection during cereal crop breeding[34,35]. *ZmPHYB2* (GRMZM2G092174) and two phytochrome-interacting factors, *ZmPIF4* (GRMZM5G865967) and *ZmPIF3.3* (GRMZM2G062541; ref. [36]) were identified as selection candidates in either the US or Chinese breeding processes. We identified six nonsynonymous SNP variants in the coding region of *ZmPIF3.3* that have MAF > 0.05. Haplotype analysis revealed that five major haplotypes formed by these six SNPs were associated with EH and that frequencies of two favorable haplotypes (Hap1 and Hap3, conferring reduced EH)

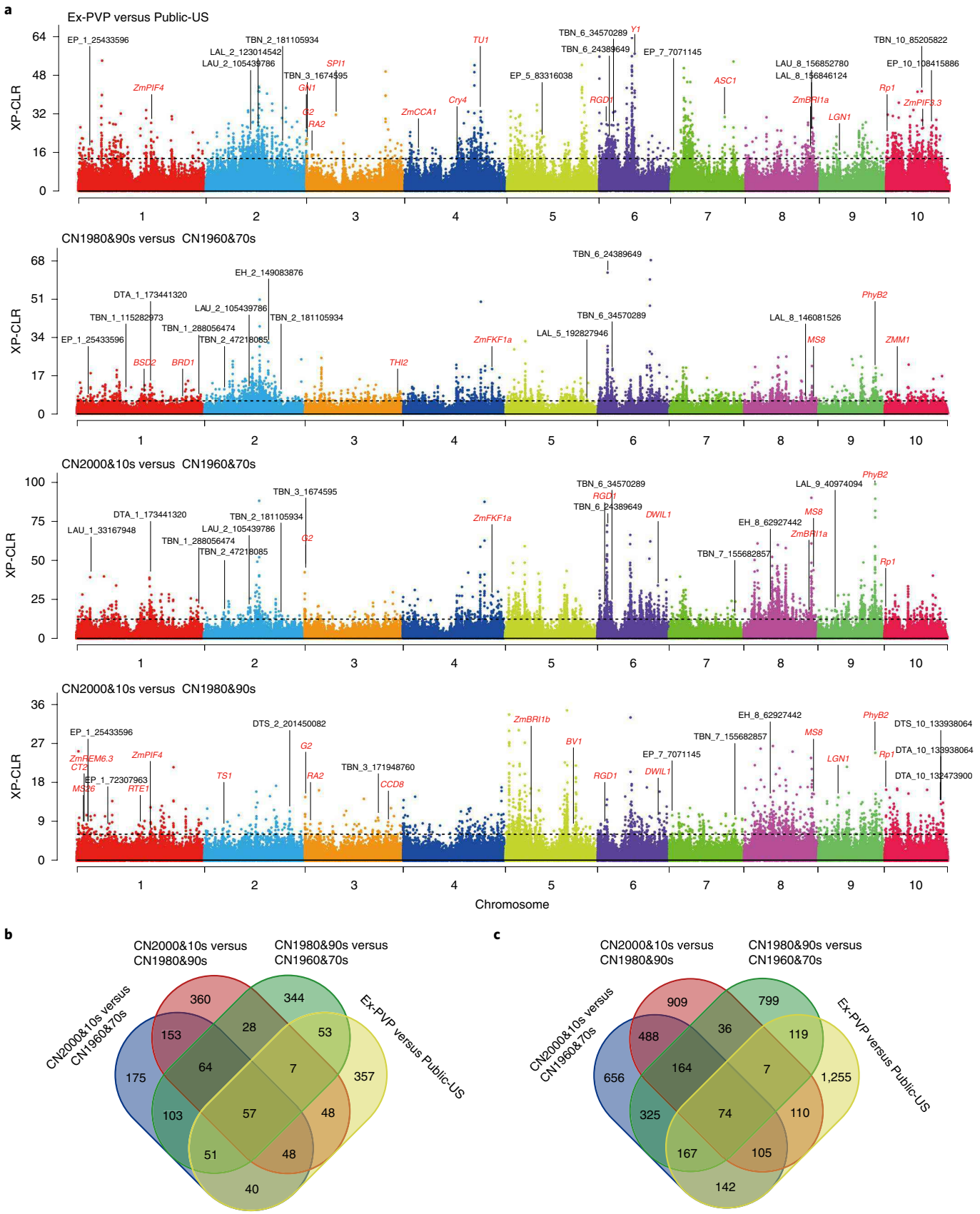


Fig. 4 | Profiling of the selective sweeps during modern maize breeding. a, Genome-wide selective signals (XP-CLR score) of different breeding eras in the United States and China. The chromosome numbers are set along the *x* axis. The horizontal black dashed lines represent the cutoffs that define statistical significance. The genes have been functionally characterized (in red) from MaizeGDB and the GWAS loci mapped in this study are marked above the selective signal peaks. **b**, Comparison of the identified selective sweeps between different breeding eras. **c**, Number of genes encompassed in the selective sweeps of the different breeding eras.

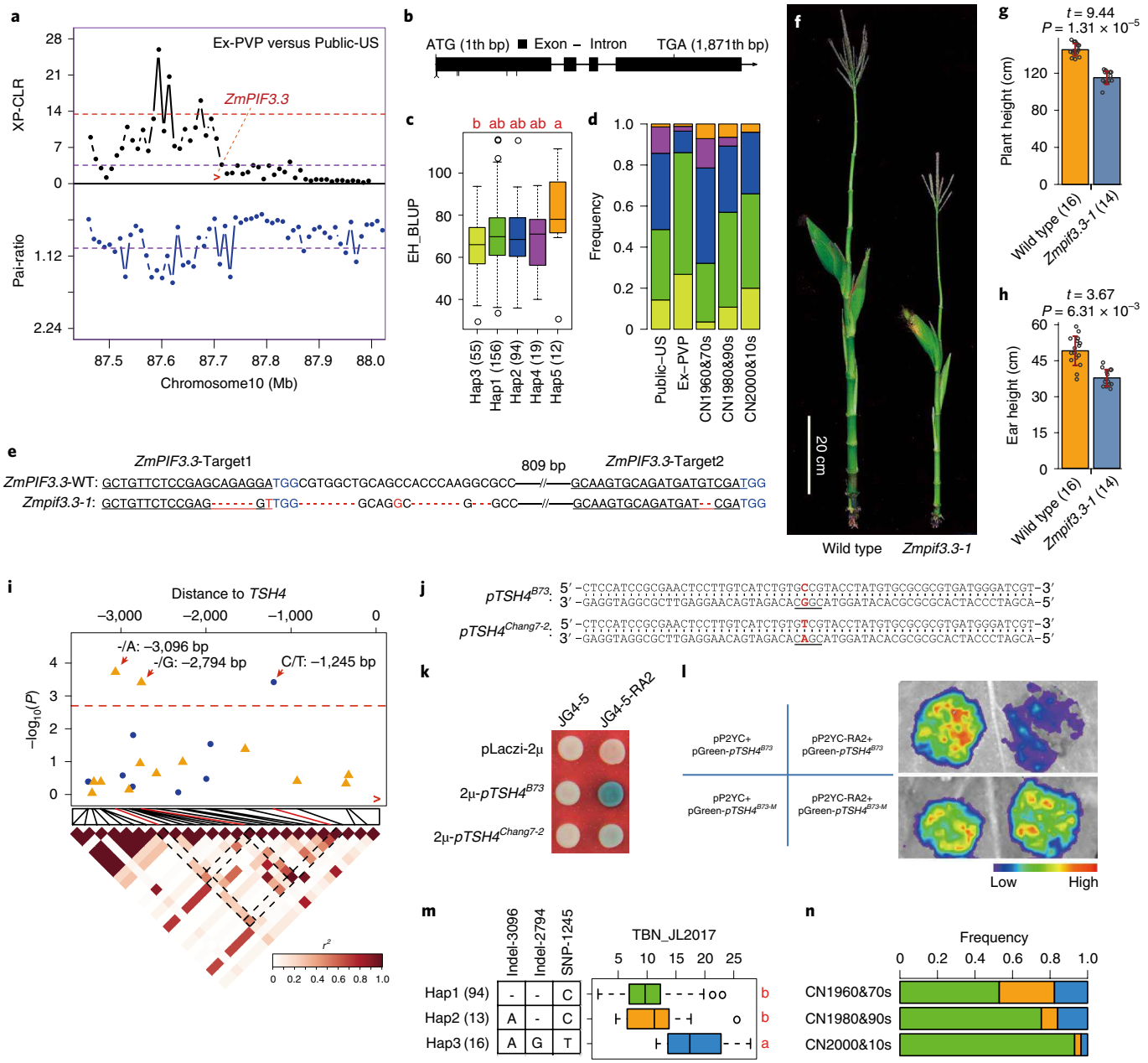


Fig. 5 | Validation of two candidate genes associated with EH and TBN. a, XP-CLR (above *x* axis) and π-ratio (below *x* axis) plot of *ZmPIF3.3*. The horizontal dashed lines represent the genome-wide cutoff, 80th quantile for XP-CLR score, and median of genome-wide $\pi_{Public\text{-}US}/\pi_{Ex\text{-}PVP}$ (top-down perspective). **b**, Gene structure and polymorphisms in *ZmPIF3.3*. **c**, Box plot for EH of five major haplotypes (more than ten inbred lines) of *ZmPIF3.3*. **d**, Haplotype frequency for *ZmPIF3.3* in different breeding eras of United States and China. **e**, Knockout of *ZmPIF3.3* by CRISPR-Cas9 system. **f**–**h**, Height profile (**f**) and statistics (**g**, plant height; **h**, ear height) of wild-type and *Zmpif3.3-1* CRISPR-knockout plants. The *P* values and *t* values of two-tailed *t*-test are shown. **i**, Manhattan plot (upper) and LD heat map (lower) for candidate association signals of TBN in the *TSH4* promoter. Association signals for SNP and indel are shown as blue dots and orange triangles, respectively. **j**, The promoter fragment of *TSH4* used for yeast one-hybrid assay. The core motifs of 5′-CGGC-3′ are shown as underscored letters. The SNP-1245 is shown in red. **k**, Yeast one-hybrid assay shows that RA2 directly binds to the $TSH4^{B73}$ promoter fragment but not that of $TSH4^{Chang7\text{-}2}$. **l**, Luciferase activity assay shows that coexpression of *RA2* effectively inhibits the *LUC* reporter gene driven by the $TSH4^{B73}$ promoter but not the $TSH4^{B73\text{-}M}$ promoter. **m**, Box plot for TBN of different haplotypes of the *TSH4* promoter (inbred lines number is shown in parenthesis). Different letters on the right indicate significant differences ($P < 0.05$, Bonferroni correction). **n**, Frequency changes of *TSH4* promoter haplotypes in different breeding eras of China.

simultaneously increased during modern maize breeding in both the United States and China (Fig. 5a–d and Supplementary Table 18). Additionally, knocking out *ZmPIF3.3* caused significantly reduced EH ($P = 6.31 \times 10^{-3}$; Fig. 5e–h and Extended Data Fig. 9), thus confirming a critical role for *ZmPIF3.3* in regulating EH and as a selective target.

SQUAMOSA Promoter Binding Protein-like (*SPL*) genes play critical roles in regulating various morphological traits in maize[37]. Notably, our GWAS identified a SNP (chr7_133305039, $P = 6.83 \times 10^{-8}$) and an indel (indel-3096, $P = 1.86 \times 10^{-5}$) that were significantly associated with TBN (Extended Data Fig. 10). The SNP and the indel are located ~98.852 and ~3.096 kb upstream of a *SPL* gene, *TSH4*, respectively. Consistent with an earlier report[38], *tsh4* knockout mutants generated via CRISPR–Cas9 technology showed dramatically reduced TBN (Extended Data Fig. 9). To identify the potential

causal variations, we performed promoter sequencing analysis of 123 inbred lines. Candidate associate mapping identified two new TBN-associated variants (indel-2794 and SNP-1245, located 2794 and 1245 bp upstream of *TSH4*, respectively) (Fig. 5i). Notably, SNP-1245 is located in a core binding motif (5′-CGGC-3′) for LATERAL ORGAN BOUNDARIES (LOB) transcription factors[39]. Previous studies reported that *ROMASA2* (*RA2*) encodes a LOB domain transcription factor playing an important role in determining stem cell fate in the maize tassel branch meristems[40]. Yeast one-hybrid assay showed that RA2 could directly bind to the *TSH4* promoter fragment containing wild-type CGGC motif (B73 type, $TSH4^{B73}$) but not the *TSH4* promoter fragment containing mutated CGGC motif (Chang7-2 type, $TSH4^{Chang7\text{-}2}$) (Fig. 5j,k). Luciferase activity assay showed that coexpression of *RA2* effectively inhibited the *LUC* reporter gene driven by the $TSH4^{B73}$ promoter but not the $TSH4^{B73\text{-}M}$ promoter (in which the CGGC motif was mutated to the Chang7-2 type, Fig. 5l). These results are consistent with the observations that *RA2* and *TSH4* have complementary expression domains in the male inflorescence[38]. Moreover, haplotype analysis based on the three associated variants (indel-3096, indel-2794 and SNP-1245) revealed that the frequency of the favorable haplotype (Hap1, conferring reduced TBN) was significantly increased during Chinese maize breeding ($P < 0.05$, Bonferroni correction, Fig. 5m,n), thus validating *TSH4* as a selective target during modern maize breeding.

In sum, the results presented here provide a valuable resource for mining of superior alleles for adaptation to high-density planting. More broadly, we demonstrate how careful sampling across eras of crop breeding, combined with phenotypic association and selection scans, can lead to high-confidence candidates that can then be functionally validated using modern gene editing technology. This pipeline could easily be applied across agronomic systems for more efficient and targeted plant breeding.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; detainbred lines of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-0616-3.

Received: 22 June 2019; Accepted: 23 March 2020;
Published online: 27 April 2020

References

1. Andorf, C. et al. Technological advances in maize breeding: past, present and future. *Theor. Appl. Genet.* **132**, 817–849 (2019).
2. Duvick, D. Genetic progress in yield of United States maize (*Zea mays* L.). *Maydica* **50**, 193–202 (2005).
3. Duvick, D. The contribution of breeding to yield advances in maize (*Zea mays* L.). *Adv. Agron.* **86**, 83–145 (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. *Food and Agriculture Organization of the United Nations Agriculture Databases* (FAO, 2016); http://www.fao.org/statistics/databases/en/.
6. Wright, S. I. et al. The effects of artificial selection of the maize genome. *Science* **308**, 1310–1314 (2005).
7. Yamasaki, M. et al. A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. *Plant Cell* **17**, 2859–2872 (2005).
8. Hufford, M. B. et al. Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **44**, 808–811 (2012).
9. Unterseer, S. et al. A comprehensive study of the genomic differentiation between temperate Dent and Flint maize. *Genome Biol.* **17**, 137 (2016).
10. Swarts, K. et al. Genomic estimation of complex traits reveals ancient maize adaptation to temperate North America. *Science* **357**, 512–515 (2017).
11. Wang, L. et al. The interplay of demography and selection during maize domestication and expansion. *Genome Biol.* **18**, 215 (2017).
12. Liu, H. et al. Genomic, transcriptomic, and phenomic variation reveals the complex adaptation of modern maize breeding. *Mol. Plant* **8**, 871–884 (2015).
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. Van Heerwaarden, J., Hufford, M. B. & Rossibarra, J. Historical genomics of North American maize. *Proc. Natl Acad. Sci. USA* **109**, 12420–12425 (2012).
15. Jiao, Y. et al. Genome-wide genetic changes during modern breeding of maize. *Nat. Genet.* **44**, 812–815 (2012).
16. 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).
17. Brekke, B., Edwards, J. & Knapp, A. Selection and adaptation to high plant density in the Iowa Stiff Stalk Synthetic maize (*Zea mays* L.) population. *Crop Sci.* **51**, 1965–1972 (2011).
18. Zhang, R. et al. Patterns of genomic variation in Chinese maize inbred lines and implications for genetic improvement. *Theor. Appl. Genet.* **131**, 1207–1221 (2018).
19. Grotewold, E., Drummond, B. J., Bowen, B. & Peterson, T. The myb-homologous *P* gene controls phlobaphene pigmentation in maize floral organs by directly activating a flavonoid biosynthetic gene subset. *Cell* **76**, 543–553 (1994).
20. Buckner, B., Kelson, T. L. & Robertson, D. S. Cloning of the *y1* locus of maize, a gene involved in the biosynthesis of carotenoids. *Plant Cell* **2**, 867–876 (1990).
21. Tan, B. C. et al. Structure and origin of the *White Cap* locus and its role in evolution of grain color in maize. *Genetics* **206**, 135–150 (2017).
22. 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).
23. Xu, B. et al. Overexpression of *AtLOV1* in switchgrass alters plant architecture, lignin content, and flowering time. *PLoS ONE* **7**, e47399 (2012).
24. Moreno, M. A., Harper, L. C., Krueger, R. W., Dellaporta, S. L. & Freeling, M. *liguleless1* encodes a nuclear-localized protein required for induction of ligules and auricles during maize leaf organogenesis. *Genes Dev.* **11**, 616–628 (1997).
25. Tian, F. et al. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* **43**, 159–162 (2011).
26. Rawat, R. et al. REVEILLE1, a Myb-like transcription factor, integrates the circadian clock and auxin pathways. *Proc. Natl Acad. Sci. USA* **106**, 16883–16888 (2009).
27. Fujiwara, S. et al. Circadian clock proteins LHY and CCA1 regulate SVP protein accumulation to control flowering in *Arabidopsis*. *Plant Cell* **20**, 2960–2971 (2008).
28. Hwang, I. & Sheen, J. Two-component circuitry in *Arabidopsis* cytokinin signal transduction. *Nature* **413**, 383–389 (2001).
29. Du, Y. et al. *UNBRANCHED3* regulates branching by modulating cytokinin biosynthesis and signaling in maize and rice. *New Phytol.* **214**, 721–733 (2017).
30. Stephan, W. Signatures of positive selection: from selective sweeps at individual loci to subtle allele frequency changes in polygenic adaptation. *Mol. Ecol.* **25**, 79–88 (2016).
31. Ross-Ibarra, J., Morrell, P. L. & Gaut, B. S. Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proc. Natl Acad. Sci. USA* **104**, 8641–8648 (2007).
32. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402 (2010).
33. Ueguchi-Tanaka, M. et al. *GIBBERELLIN INSENSITIVE DWARF1* encodes a soluble receptor for gibberellin. *Nature* **437**, 693–698 (2005).
34. Sawers, R. J., Sheehan, M. J. & Brutnell, T. P. Cereal phytochromes: targets of selection, targets for manipulation? *Trends Plant Sci.* **10**, 138–143 (2005).
35. Sheehan, M. J., Kennedy, L. M., Costich, D. E. & Brutnell, T. P. Subfunctionalization of *PhyB1* and *PhyB2* in the control of seedling and mature plant traits in maize. *Plant J.* **49**, 338–353 (2007).
36. Wu, G. et al. Characterization of maize Phytochrome-Interacting Factors in light signaling and photomorphogenesis. *Plant Physiol.* **181**, 789–803 (2019).
37. Wei, H., Zhao, Y., Xie, Y. & Wang, H. Exploiting *SPL* genes to improve maize plant architecture tailored for high density planting. *J. Exp. Bot.* **ery258**, 1–14 (2018).
38. Chuck, G., Whipple, C., Jackson, D. & Hake, S. The maize SBP-box transcription factor encoded by *tasselsheath4* regulates bract development and the establishment of meristem boundaries. *Development* **137**, 1243–1250 (2010).
39. Husbands, A., Bell, E. M., Shuai, B., Smith, H. M. & Springer, P. S. LATERAL ORGAN BOUNDARIES defines a new family of DNA-binding transcription factors and can interact with specific bHLH proteins. *Nucleic Acids Res.* **35**, 6663–6671 (2007).
40. Bortiri, E. et al. *ramosa2* encodes a LATERAL ORGAN BOUNDARY domain protein that determines the fate of stem cells in branch meristems of maize. *Plant Cell* **18**, 574–585 (2006).

Methods

Plant materials and phenotypic measurement. In total, 350 elite inbred lines of maize were collected in our study (Supplementary Table 1). The US inbred lines were selected on the basis of published literature or information provided in the Germplasm Resources Information Network website (https://www.ars-grin.gov). The elite Chinese inbred lines were selected on the basis of publically available pedigrees, registration information of the hybrids and personal communications with maize breeders.

The 350 inbred lines were planted and phenotyped across four environments for 2 consecutive years: the city of Langfang in Hebei province in 2016 (LF2016) and 2017 (LF2017), Ledong County in Hainan province in 2016 (HN2016) and the city of Gongzhuling in Jilin province in 2017 (JL2017) in China. A randomized complete block design was used in all four trials. In the LF2016 and HN2016 trials, we had one replicate, while in the LF2017 and JL2017 trials we had two replications for each inbred line. The row and column spacing was respectively set to 0.600 and 0.278 m. At least five plants in the middle of the plot were selected for phenotyping (Supplementary Table 2).

The inbred lines A71 (Public-US inbred lines with large leaf angle) and 2MA22 (Ex-PVP inbred lines with erect leaf angle) were used for expression analysis of *ZmNAC16*, while H49 (Public-US inbred line with large leaf angle) and 2MA22 were used for expression analysis of *ZmSBP18* (A71 and 2MA22 share the same haplotype at this locus). The leaf collars of expanded and unexpanded leaves from V7 seedlings were collected for RNA extraction (three to four sampling replicates for each line and each replicate consists of collar tissues from three independent plants).

Statistical analysis. The trait values of 15 traits across all trials were fit by a linear mixed model in R with the lme4 package[41] to obtain a best linear unbiased predictor (BLUP) value as follows:

$$Y_{ij} = \mu + \mathrm{Line}_i + \mathrm{Env}_j + (\mathrm{Line}\times\mathrm{Env})_{ij} + (\mathrm{Env}\times\mathrm{Rep})_{jn} + \mathrm{error}_{ijn}$$

where μ is the mean, Line_i is the genotype effect of the i-th inbred, Env_j is the effect of the j-th environment, $(\mathrm{Line}\times\mathrm{Env})_{ij}$ is the genotype–environment interaction and $(\mathrm{Env}\times\mathrm{Rep})_{jn}$ is the environment–replication interaction, error_{ijn} is the error of the j-th environment and the n-th replication and items were set to random. Multiple comparisons of the trait values were conducted by the least significant difference (LSD) method with R package agricolae (https://cran.r-project.org/web/packages/agricolae/). A Bonferroni corrected $P<0.05$ was used to define the statistical significance of multiple comparisons.

DNA isolation and sequencing. Six inbred lines (Mo17, Zheng58, Chang7-2, 478, 5003 and 8112) have been previously sequenced to more than 20× (ref. [15]). The young seedlings of 344 inbred lines were collected and their genomic DNA was extracted with the cetyl trimethylammonium bromide (CTAB) method. The genome sequencing libraries were sequenced with Illumina X-ten sequencer, yielding a total of ~69.3×10^9 150-bp paired-end reads (~10.4 terabases, depth range from 10.0× to 28.9×; Supplementary Table 1).

Variant calling, evaluation and annotation. The quality of the short sequencing reads was first evaluated with FastQC (v.0.10.1) and controlled with Trimmomatic (v.0.36). We used BWA mem[42] (v.0.7.13) to map the remaining clean reads against the B73 reference genome (RefGen_v3) with default parameters. The mapping results were processed with SAMtools[43] (v.1.3.1). GATK[44] (HaplotypeCaller function; v.3.5-0-g36282e4) was used to call the raw variants by following the best-practice workflows. To obtain high-quality variants, we retained variants with: QD > 2.0, FS < 60.0, MQ > 20.0, MQRankSum > −12.5 and ReadPosRankSum > −8.0. Further, we removed the following potential low-quality variants: (1) missing rate >80%, (2) frequency of heterozygous genotype >5% or more than twice the minor homozygous allele frequency and (3) deviated from Hardy–Weinberg expectation as proposed in the GATK (ExcessHet < 1×10^{-5}). After filtering, the genomic heterozygous rates of all inbred lines were 1.21% on average, similar to the results from maize HapMap 1 (ref. [45]).

To evaluate the quality of variants, we compared our SNPs to those of maize HapMap 3 (ref. [46]). Among the seven deeply sequenced (>10×) inbred lines in HapMap 3, the SNP concordance varied from 93.6% to 98.8%, with a mean of 96.8% (Supplementary Table 4). We also resequenced the reference genome inbred line B73 and its variants relative to the B73 reference genome were called and filtered as described above. We identified three types of variants in the resequenced B73: (1) the sequenced B73 with the same genotype as the reference one (99.51% of all SNPs), (2) the resequenced B73 with missing genotype (109,681, 0.43% of all SNPs) and (3) the resequenced B73 with nonmissing variants regarding reference B73 (14,904, 0.06% of all SNPs). We found that more than 64% of the missing genotypes were low mapping quality reads (multiple mapping) that were filtered out in our pipeline and 31% were due to sequencing gaps, whereas the remaining 5% were supported by less than 30 reads (thus filtered out). Overall, our SNP accuracy was estimated to be >99.5%.

SNP annotation was performed according to the gene model (Zea_mays.AGPv3.31.chr.gff3.gz) generated on the basis of the B73 reference genome (GCA_000005005.5_B73_RefGen_v3) by using SnpEff (v.4.3a). We found that 7,469,468 SNPs are in genic regions and 17,851,196 SNPs are in intergenic regions. In coding regions, the ratio of nonsynonymous-to-synonymous substitutions is 1.086 (Supplementary Table 3), which is larger than the value of 0.696 reported in maize HapMap 2 (a panel of wild relatives, landraces and diverse inbred lines)[47] but is similar to an earlier report using an elite inbred lines panel[15].

Population genetic analyses. To evaluate the nucleotide diversity of inbred lines from different eras, we calculated θ_π in a 10-kb nonoverlapping window using the libsequence C++ library[48] and inhouse Perl scripts as reported[15]. The average θ_π of the genome-wide 10-kb windows was selected to represent the nucleotide diversity of inbred lines of every era.

LD (calculated as r^2) in the study was calculated using SNPs with MAF > 0.05 and missing rate <0.5 by PLINK[49] (v.1.90b3.42) with the following parameters: --ld-window-r2 0 --ld-window 99999 --ld-window-kb 500. The genome-wide average r^2 between two SNPs within 500-kb windows was calculated and the distance of LD decay was represented as the physical distance over which r^2 drops to 0.2.

We first conducted principle component analysis (PCA) with Eigensoft[50] using all 25,320,665 SNPs and selected the first three principle components (PCs, explained 19.16% of the genotypic variation) to represent the population structure of inbred lines. To further confirm the result, we also conducted a model-based method implemented in the ADMIXTURE tool[51]. We determined the number of ancestry populations of inbred lines (K) with a fivefold cross-validation approach implemented in the tool. We observed that when $K=4$, the cross-validation error was sharply convergent, suggesting that $K=4$ is a reasonable number for the ancestries of these inbred lines. On the basis of $K=4$, the inbred lines were grouped into a mixed group and four groups corresponding to SS, NSS, HZS and IDT. Each inbred line was assigned to one of the four groups if the group contributed more than 60% of its genome. Otherwise it was assigned to the mixed group.

Selective sweep detection. A composite likelihood approach (XP-CLR) was used to scan for the genome-wide selective sweeps[8,32]. We used the earlier breeding era as a reference and the later one as a query to identify the potential breeding sweeps (CN1980&90s versus CN1960&70s, CN2000&10s versus CN1980&90s, CN2000&10s versus CN1960&70s and Ex-PVP versus Public-US). The reported high-density genetic map constructed from the nested association mapping (NAM) populations of Chinese inbred lines[52] was used, and the genetic distance between adjacent markers was interpolated according to their physical distance in the genetic map. Selective sweeps were identified following a previously described procedure[8]. In brief, we scanned the selective sweeps with a step of 100 bp and a sliding window of 0.05 cm and grouped nonoverlapping 10-kb windows across the genome into features, considering the nonindependence of genome regions. The XP-CLR score and selection coefficient of every feature were obtained by averaging over all original 100-bp steps included in the feature. We then selected the top 20% of features as the putatively selected features. To control the potential effect of asymmetrical population structure, we excluded the IDT germplasm in the selective sweep analysis as it was only used in the most recent breeding eras of the United States and China. To ensure that interpopulation selective sweeps did not overlap and confound interera selective sweep, we identified selective sweeps in all subpopulations among different breeding stages with the same pipeline and removed the putatively selected features of all inbred lines that were not supported by the selective features of any involved subpopulation. To minimize the sample size effect, Chinese inbred lines were divided into two groups, CN1960-80s and CN1990-2010s, during subpopulation selective sweeps analysis. We selected the top 10% of the controlled putatively selected features as the potential selective sweeps. Next, we calculated nucleotide diversity (π) for 10-kb nonoverlapping windows in inbred lines of each breeding era and investigated the π ratio for all windows. Based on the genome-wide π ratio, we removed the selective sweeps with ratios lower than the median of the genome-wide values as implemented in Hufford et al.[8]. Lastly, we merged adjacent selective sweeps with distance less than 10 kb and selective sweeps from different comparisons with at least 50% overlap were deemed the same selective sweeps.

To further confirm the selective sweeps, we also investigated the interera *Fst* among the inbred lines by a slide window approach with a window size of 100 kb and a step of 10 kb using Vcftools. To validate that the large linkage block on 25.5 Mb of chromosome 1 (associated with EP) resulted from selection, we investigated the XP-EHH statistics (detection of selective sweeps based on extended haplotype blocks) for every SNP using Selscan program[53].

To identify candidate genes for all selective sweeps, we first included all annotated genes that were located directly in the sweeps. For sweeps without any gene, the gene closest (<33 kb) to the XP-CLR peak of the selective sweep was deemed as the potential candidate for the sweep, as described[8].

For gene ontology enrichment analysis, we performed gene ontology annotation for all maize annotated genes with PANNZER2 (ref. [54]) with default settings. The Fisher's exact test was then used to identify potentially significantly enriched gene ontology terms ($P<0.05$).

GWAS of plant morphological traits. We selected 11,622,737 SNPs (MAF > 0.05 and missing rate <50%) to perform GWAS of all traits. The missing genotypes

were imputed with Beagle (v.4.1) with default parameters. The GWAS was conducted with a linear mixed model that was implemented in the EMMAX package[55]. We performed GWAS using both the BLUP and single trial values for all traits. To determine the genome-wide significant cutoff for GWAS results, we estimated the number of genome-wide effective SNPs by pruning SNPs within 500 bp and with a $r^2 \geq 0.2$ by a slide window approach with a window size of 500 bp and step of 100 bp using PLINK. After pruning, the number of effective SNPs was determined to be 193,902. We then selected 1×10^{-6} (Benjamini–Hochberg FDR < 0.05) as the genome-wide significant cutoff. We determined significant signals with the following two criteria: (1) the *P* values of the signals for BLUP values were $<1 \times 10^{-6}$ or (2) the *P* values of signals were consistently lower than 1×10^{-5} for at least two environmental trials. For adjacent GWAS loci (<500 kb), loci independence was determined by pairwise linkage analysis of significant SNPs (if $r^2 < 0.5$, they were declared independent). The confidence intervals of the GWAS loci were determined by local LD block analysis where pairwise r^2 of the SNPs with $P < 1 \times 10^{-3}$ should be >0.3. Genes located directly in or within 33 kb (genome-wide average distance of LD decay to $r^2 = 0.2$) around the confidence interval were selected as the candidate genes for the GWAS loci. Candidate gene-based association analysis was conducted using the Mixed Linear Model (MLM) method in Tassel5 (ref. [56]) (v.5.2.22) with small indels (<10 bp) located around the confidence interval of the GWAS loci.

The identified GWAS loci were compared to previously identified GWAS QTNs for 15 morphological traits from the NAM, CN-NAM, 508 diverse inbred lines (AM508) and ten Recombinant Inbred Line (RIL) populations (Supplementary Table 7). Since the positions of these reported QTNs were based on the B73_RefGen_v1 and B73_RefGen_v2, their corresponding positions on the B73_RefGen_v3 reference genome were determined by BLASTing their 200 bp flanking sequences against the B73_RefGen_v3 reference.

Candidate association mapping of *TSH4* for TBN. Promoter sequences (3.4 kb) of *TSH4* of 123 inbred lines (115 are from China) were amplified by PCR using the primer pairs p7588-1 and tsh4-4 (or tsh4-7; Supplementary Table 19), followed by Sanger sequencing. Candidate association was performed using the MLM method in Tassel5 (v.5.2.22). Three variants were found significantly associated with TBN variation in the *TSH4* promoter (Bonferroni FDR < 0.05). The top signal (−3,096 bp) was strongly associated with indel-2794 and SNP-1245 (r^2 both equal to 0.48). The indel-2794 and SNP-1245 are in complete linkage ($r^2 = 1$).

qRT–PCR analysis of candidate genes. RNA was extracted using TRIzol reagent (Invitrogen). Complementary DNA was synthesized using the FastQuant RT Kit (Tiangen, catalog no. KR108-02). Quantitative reverse transcription PCR was performed using SuperReal PreMix Plus (Tiangen, FP205-2) and a QuanStudio 3 Real Time PCR System cycler (Applied Biosystems). *Tubllin5* was used as the internal control. The primers used for qRT–PCR are listed in Supplementary Table 19.

Knockout of *ZmPIF3.3* and *TSH4* by CRISPR–Cas9 system. The CRISPR–Cas9 constructs for *ZmPIF3.3* and *TSH4* were generated using a previously described vector[57]. The multiple target sequences designed for these genes are shown in Fig. 5. All constructs were introduced into the *Agrobacterium* strain EHA105 and transformed into the immature embryo of the maize inbred line ZC01 through *Agrobacterium*-mediated transformation.

The target regions of *ZmPIF3.3* and *TSH4* were amplified from ZC01 and corresponding transgenic lines and sequenced to identify the mutations. For *ZmPIF3.3*, we obtained three independent homozygous knockout lines named *Zmpif3.3-1*, *Zmpif3.3-2* and *Zmpif3.3-3*. For *TSH4*, we obtained two independent homozygous knockout lines named *tsh4-1* and *tsh4-2* (Extended Data Fig. 9).

The phenotypes of these mutants were investigated under normal field planting conditions, together with their wild-type ZC01. The *Zmpif3.3-1* mutant was planted in Ledong County in Hainan province in 2017, while the *Zmpif3.3-2* and *Zmpif3.3-3* mutants were planted in the same location in 2018. The *tsh4-1* and *tsh4-2* mutants were planted in Langfang in Hebei province in 2017. Each mutant plot was planted in replicate with a neighboring wild-type control plot. Two replicates were used for these phenotyping trials. The row and column spacing were set to 0.60 and 0.25 m, respectively. The traits of PH, EH and TBN were measured as described in Supplementary Table 2.

Yeast one-hybrid assay and luciferase activity assay. For yeast one-hybrid assay, the coding region of *RA2* was PCR amplified from cDNA of inbred line B73 and ligated into the pJG4-5 vector to generate JG4-5-RA2. The ~300-bp promoter sequences around SNP-1245 in B73 and Chang7-2 are identical except the SNP-1245 variation. Thus the promoter fragment (60 bp; Fig. 5j) including the SNP-1245 was amplified from B73 and Chang7-2 respectively, using the primer pair S1245 (Supplementary Table 19), then ligated into the pLaczi2μ vector[58] to produce 2μ-$pTSH4^{B73}$ and 2μ-$pTSH^{Chang7\text{-}2}$. Yeast one-hybrid assay was conducted following a previously described protocol[59].

For luciferase activity assay, the coding region of *RA2* was cloned into the pP2YC vector to generate pP2YC-RA2. The *TSH4* promoter of B73 was amplified using the primer pair p7588-1 (Supplementary Table 19). The *TSH4* promoter with mutated 5′-CGGC-3′ motif, $pTSH4^{B73\text{-}M}$ (in which the 5′-CGGC-3′ motif was changed to Chang7-2 type) was generated by PCR using primer pairs designed according to Agilent Technologies (http://www.genomics.agilent.com). These two promoters were subsequently ligated into the plasmid pGreenII0800-LUC (Biovector) to generate pGreen-$pTSH4^{B73}$ and pGreen-$pTSH4^{B73\text{-}M}$. The pP2YC-RA2, pGreen-$pTSH4^{B73}$ or pGreen-$pTSH4^{B73\text{-}M}$ constructs were introduced into the *Agrobacterium* strain EHA105, and the cells containing pP2YC-RA2 or pGreen-fused construct were coinjected into *Nicotiana benthamiana* leaves. The infiltrated plants were incubated at 25 °C in darkness for 1 d and then in light for 1 d, then the activity of the luciferase reporter gene was examined using the NightSHADE LB985 Plant Imaging System (Berthold).

Analysis of favorable allele changes during breeding. The allele types associated with reduced ear height, more erect leaf, reduced tassel branch number or accelerating flowering were deemed to be the favorable alleles. We investigated the frequency changes of favorable alleles during breeding using the SNP set used in the GWAS. To exclude the effect of random effect and population structure, we conducted 300× permutation tests for each trait. For each permutation, the original phenotype data were reshuffled followed by GWAS analysis as described above. All permuted QTNs were used to generate a null distribution of the favorable alleles during the different breeding stages. The favorable allele frequency changes in real data were compared to their null distribution to test whether they resulted from random (like population structure) or artificial selection.

Estimate of gene flow between US and Chinese inbred lines. To estimate the degree of gene flow between the US and Chinese inbred line populations, we analyzed the extent of genetic introgression from the US inbred lines to the Chinese inbred lines (Public-US to CN1980&90s, Public-US to CN2000&10s, Ex-PVP to CN1980&90s and Ex-PVP to CN2000&10s) using the four taxa approach[60], which calculates the excessively shared derived variants between two taxa (ABBA–BABA statistic, also known as f_d statistic). We selected 65 tropical and subtropical maize inbred lines from the maize HapMap 3 (ref. [46]) as the outgroup, the inbred lines from the breeding era of CN1960&70s as control and screened the potential introgressed regions with a window size of 100 kb. Windows with meaningless result ($f_d > 1$, $f_d < 0$ or with Patterson's *D* statistic < 0) were removed and these with strongest 5% of f_d value were selected as the potential introgressed regions.

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

Data availability

DNA-sequencing reads for all maize lines were deposited in the NCBI with the accession code of PRJNA609577 and BIGD (BIG Data Center in Beijing institute of Genomics) with the accession code of CRA002372. All phenotype data of 350 inbred maize lines are included in Supplementary Table 1. Source data for Figs. 1–3 and 5 and Extended Data Figs. 1, 2 and 8–10 are presented with the paper. All other reasonable requests for data and research materials are available via contacting the corresponding authors.

References

41. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using *lme4*. *J. Stat. Softw.* **67**, 1–48 (2015).
42. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
43. Li, H. et al. The sequence alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
44. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
45. Gore, M. A. et al. A first-generation haplotype map of maize. *Science* **326**, 1115–1117 (2009).
46. Bukowski, R. et al. Construction of the third-generation *Zea mays* haplotype map. *Gigascience* **7**, 1–12 (2018).
47. Chia, J. M. et al. Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.* **44**, 803–807 (2012).
48. Thornton, K. Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* **19**, 2325–2327 (2003).
49. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
50. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
51. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
52. Li, C. et al. Construction of high-quality recombination maps with low-coverage genomic sequencing for joint linkage analysis in maize. *BMC Biol.* **13**, 1–12 (2015).
53. Szpiech, Z. A. & Hernandez, R. D. Selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol. Biol. Evol.* **31**, 2824–2827 (2014).

54. Törönen, P., Medlar, A. & Holm, L. PANNZER2: a rapid functional annotation web server. *Nucleic Acids Res.* **46**, W84–W88 (2018).
55. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
56. Bradbury, P. J. et al. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
57. Zhao, Y. et al. An alternative strategy for targeted gene replacement in plants using a dual-sgRNA/Cas9 design. *Sci. Rep.* **6**, 23890 (2016).
58. Lin, R. et al. Transposase-derived transcription factors regulate light signaling in *Arabidopsis*. *Science* **318**, 1302–1305 (2007).
59. Xie, Y. et al. Phytochrome-interacting factors directly suppress *MIR156* expression to enhance shade-avoidance syndrome in *Arabidopsis*. *Nat. Commun.* **8**, 348 (2017).
60. Martin, S. H., Davey, J. W. & Jiggins, C. D. Evaluating the use of ABBA–BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* **32**, 244–257 (2014).

Acknowledgements

The work was supported by National Key R&D Program of China (grant no. 2016YFD0101001), the Major Program of Guangdong Basic and Applied Research (grant no. 2019B030302006), National Transgenic Science and Technology Program (grant no. 2019ZX08010003-002-004), National Natural Science Foundation of China (grant nos. 31801377, 31430008 and 31921004), the Agricultural Science and Technology Innovation Program, and Jilin Provincial Science and Technology Key Project (grant no. 20170204007NY).

Author contributions

B.W. and Haiyang Wang conceived and designed the research. B.W., J.L., Z.M., T.W., Y.L., Xinhai Li, Y.C., Y.X. and Hai Wang participated in germplasm collection. B.W., B.Z., G.S., X.M., Q.L., Z.Z., D.K., H. Wei and C.C. performed phenotypic measurement. Z.L., B.W., Xin Li, M.H., J.R.-I. and H.H. analyzed the data. Y.Z. performed plasmid construction and genetic transformation. B.W., G.W., H. Wu and R.S. characterized the CRISPR–Cas9 mutants. B.W. conducted gene expression analysis. B.W., Z.L., Xin Li, M.H., J.R.-I. and H.H. wrote the manuscript. Haiyang Wang, M.H. and J.R.-I. revised the manuscript.

Competing interests

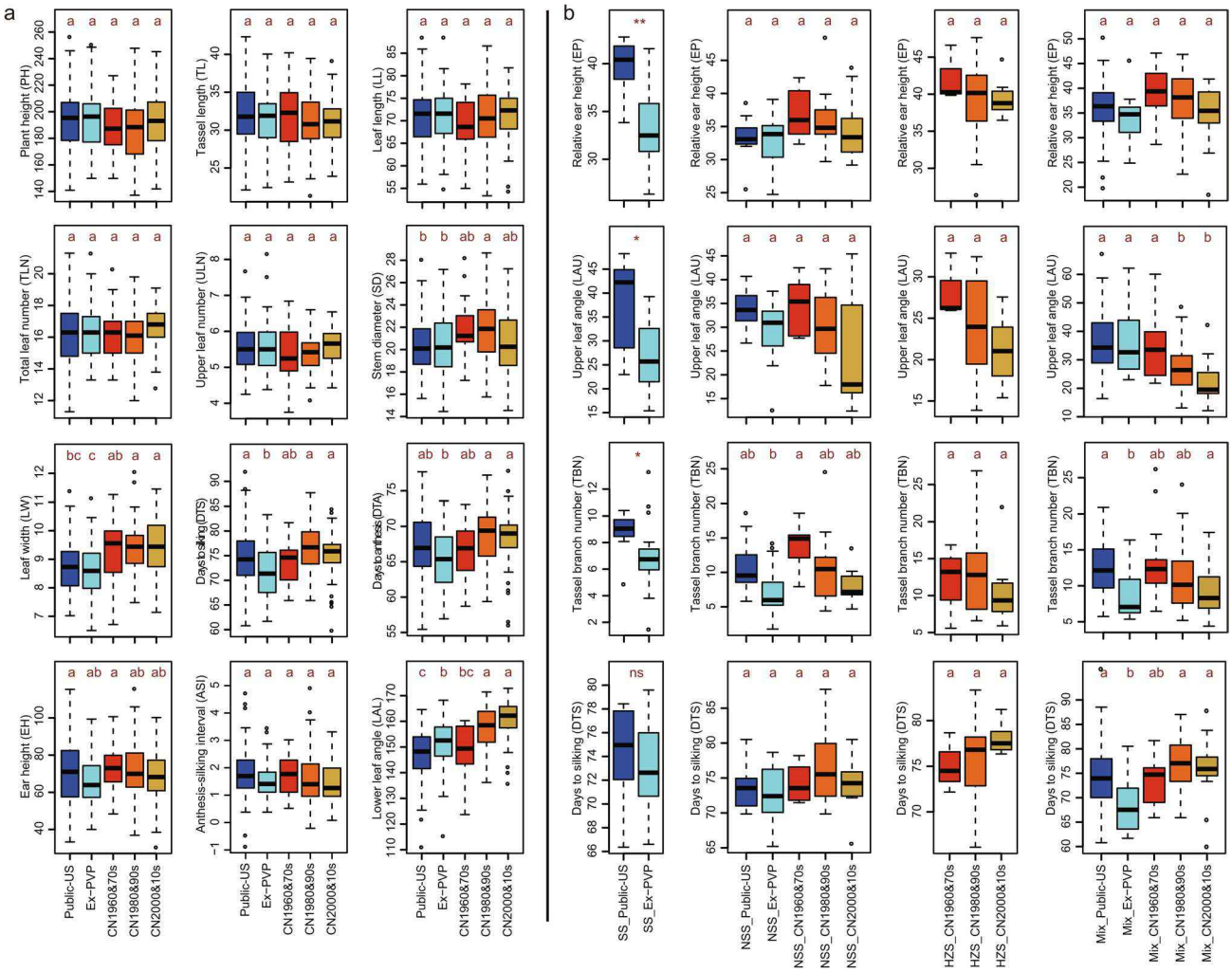
The authors declare no competing interests.

Additional information

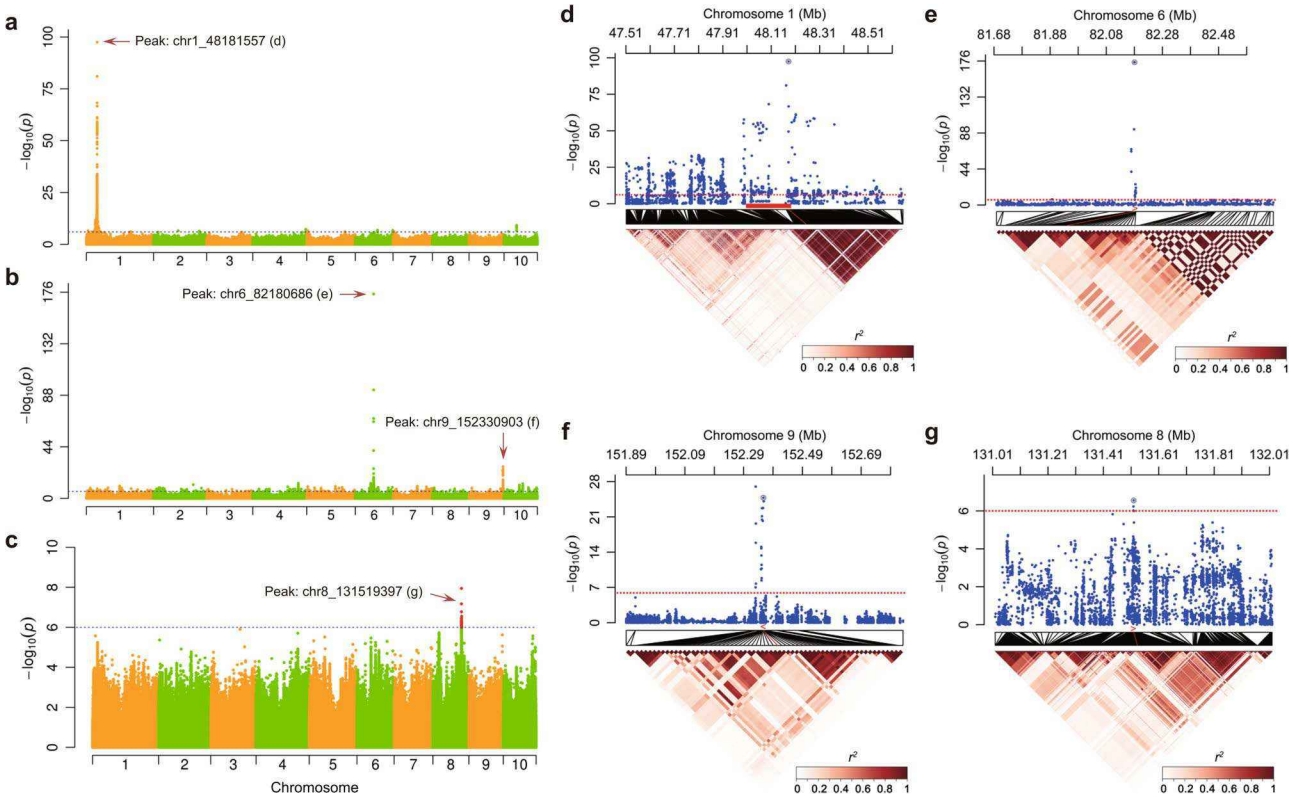
Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-0616-3.

Extended data is available for this paper at https://doi.org/10.1038/s41588-020-0616-3.

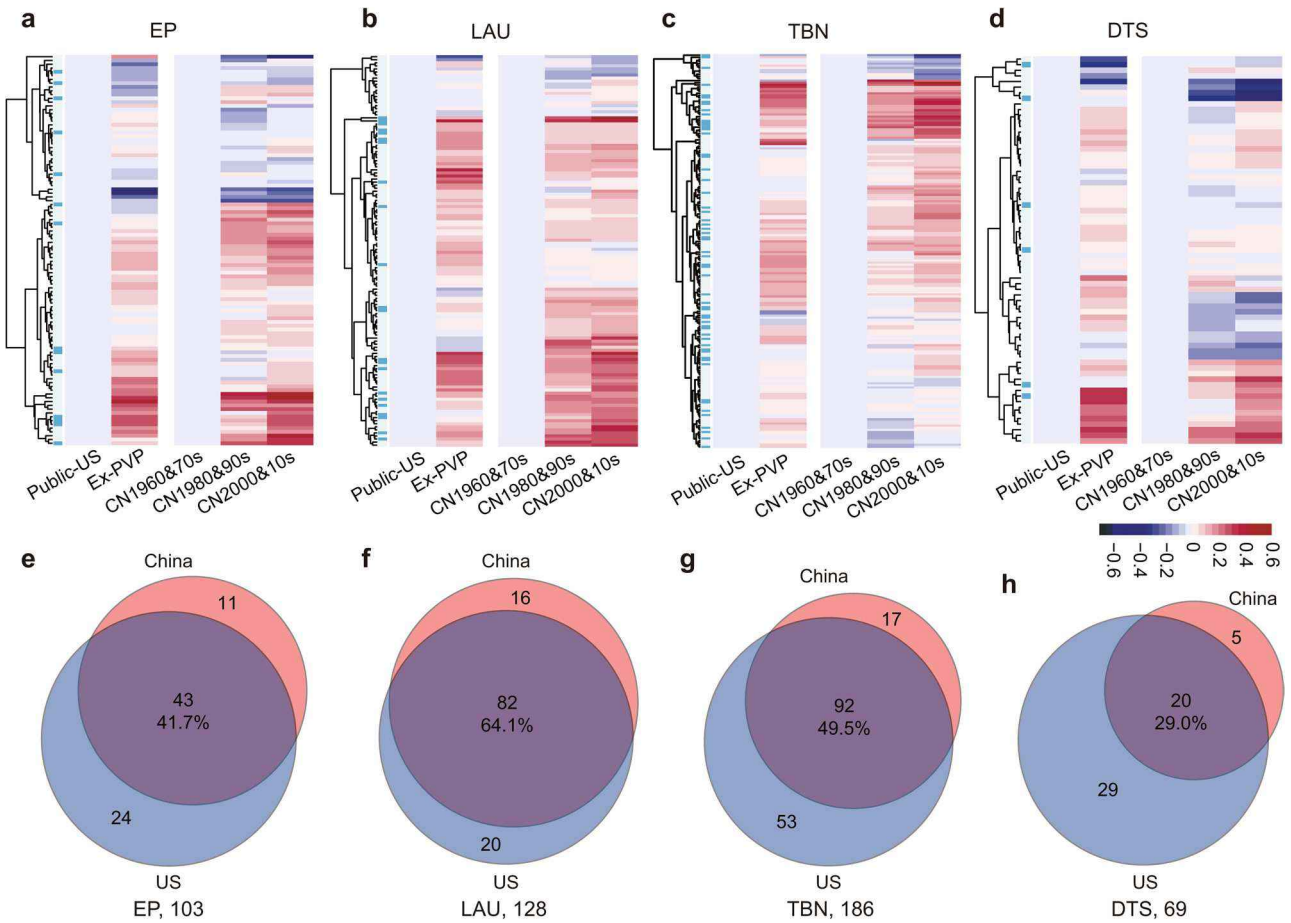
Correspondence and requests for materials should be addressed to H.H. or H.W.



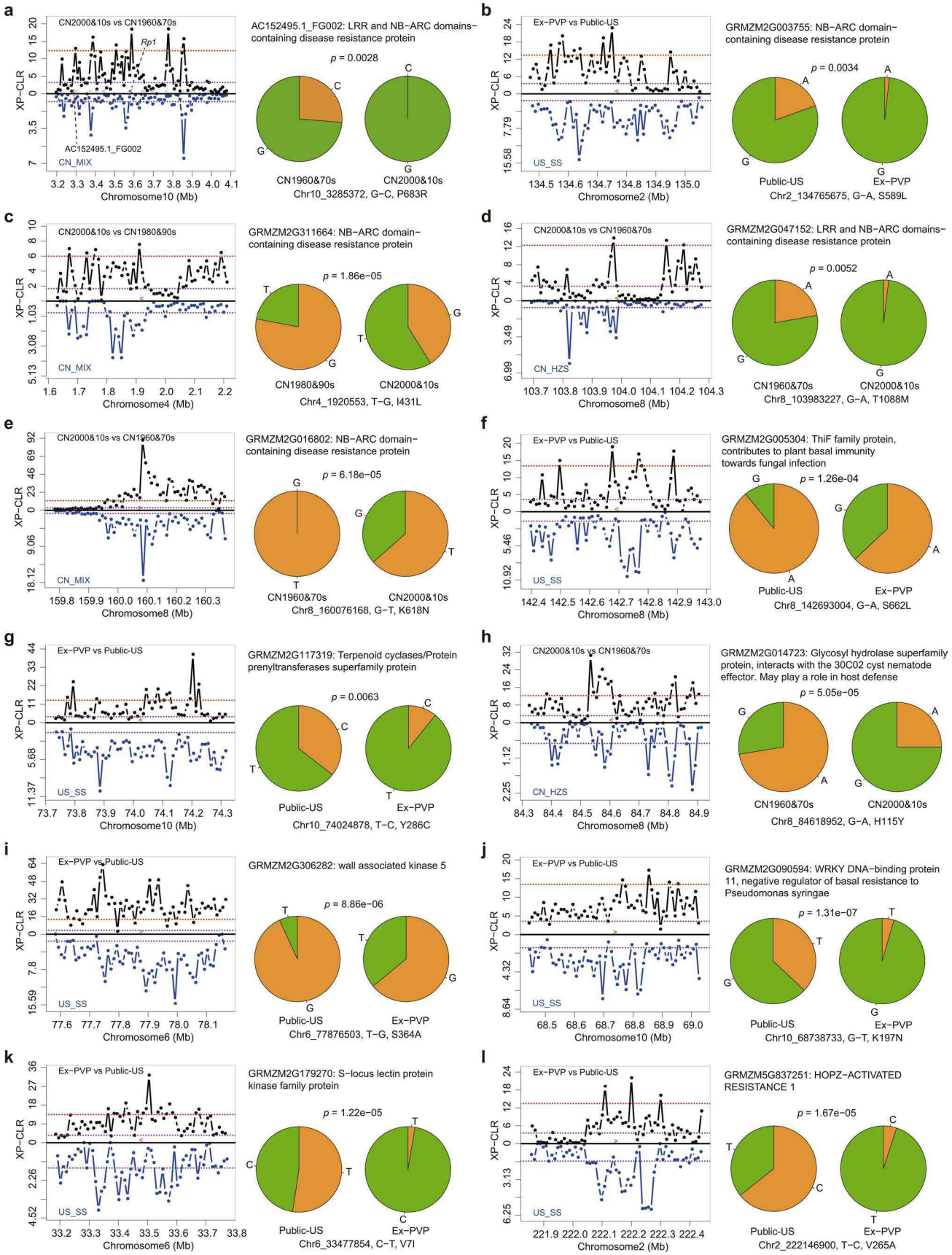
Extended Data Fig. 1 | Changes in morphological traits during maize breeding in the United States and China. a, Changes of 12 morphological traits during modern maize breeding in the United States and China. Different letters above the boxes indicate significant difference ($p < 0.05$, Bonferroni correction) in pairwise comparison. Note that days to anthesis (DTA, $p = 0.015$) and ear height (EH, $p = 0.007$) are significantly different between Public-US and Ex-PVP inbred lines as revealed by two-tailed *t*-test. **b**, Changes of four morphological traits in four subgroups (SS, NSS, HZS and Mixed) during modern maize breeding in the United States and China. Subgroups with at least 10 inbred lines in each US or Chinese era were used in the analysis. The x-axis represents the eras with prefixed sub-group names. The * or ** above the SS sub-group indicate the *t*-test results at significant level of 0.05 and 0.01, respectively.



Extended Data Fig. 2 | GWAS identification of the candidate genes for Cob Color, Kernel Color, and days to anthesis (DTA). a–**c**, Manhattan plot for Cob Color (a), Kernel Color (b) and DTA (c). **d**, *Pericarp color1* (*P1*) is associated with cob color. The peak SNP is located in the tandem repeat region of *P1*. **e**, **f**, *Yellow endosperm1* (*Y1*) and *White Cap1* (*WC1*) are associated with kernel color. The peak SNP of GWAS signal on chromosome 6 is located in the genic region of *Y1*. The second top SNP of GWAS signal on chromosome 9 is located in the genic region of *WC1*. **g**,*Vegetative to Generative Transition1* (*VGT1*) is associated with DTA. The second top SNP of GWAS signal is located within the *VGT1* region.

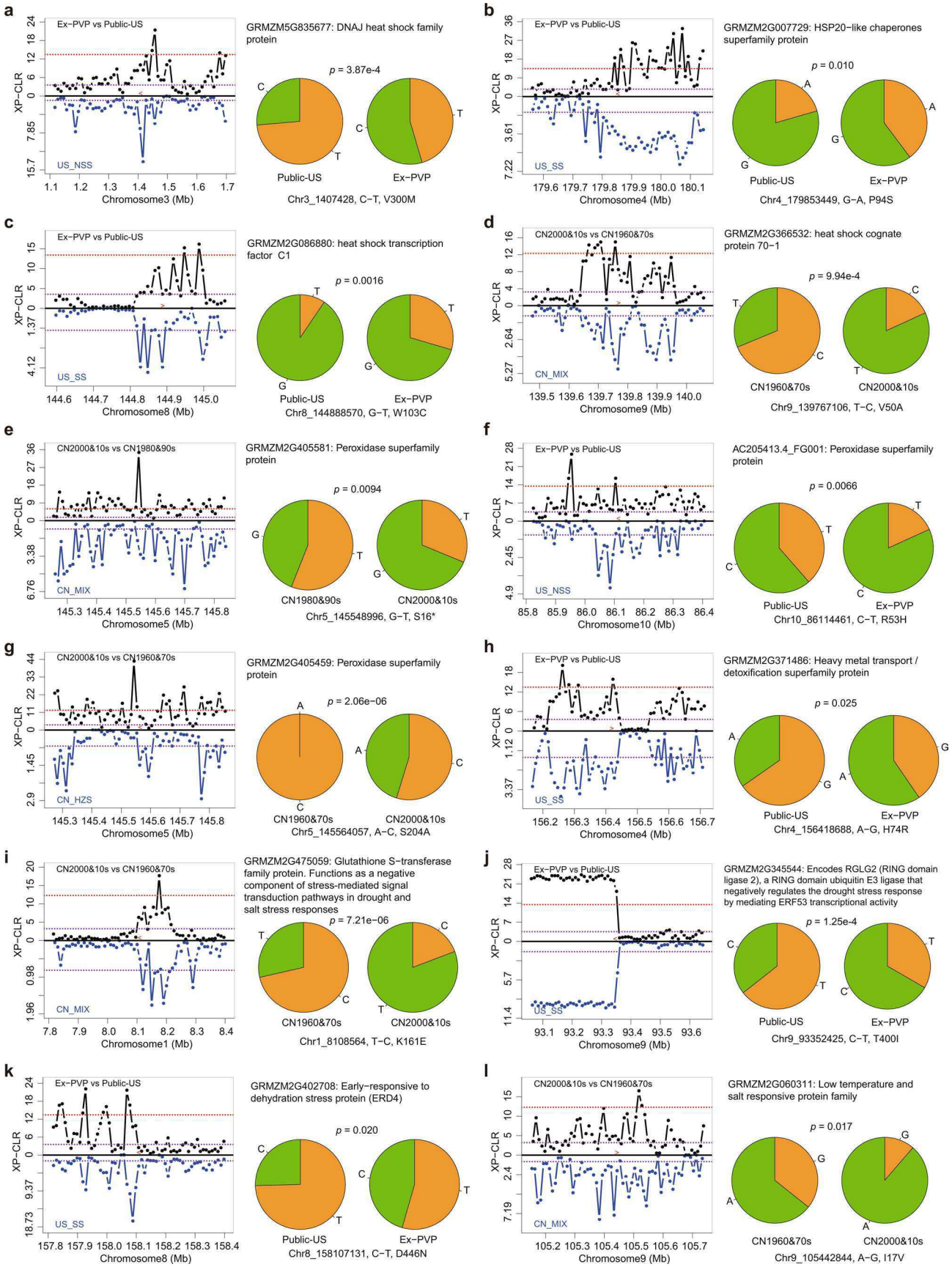


Extended Data Fig. 3 | Accumulation of favorable alleles contributes to improvement of four selected morphological traits for adaptation to high-density planting. a–d, Favorable allele frequency changing profiles of relative ear height (EP, a), upper leaf angle (LAU, b), tassel branch number (TBN, c) and days to silking (DTS, d) at QTN loci from GWAS loci during the US and Chinese inbred lines breeding process. Red indicates an increase, whereas blue indicates a decrease in the frequency of a favorable allele during breeding. Each row represents a GWAS locus, with cyan and gray colors (in the first column) mark rows representing GWAS loci obtained by the cutoff of $p < 1e\text{-}6$ and 1e-5, respectively. Later breeding stages in United States and China were compared to Public-US and CN1960&70s respectively. **e–h**, Pie plot for the numbers of GWAS loci with favorable allele frequency increased during the US and Chinese inbred lines breeding process. GWAS loci with favorable allele frequency increased during both CN1960&70s-CN1980&90s and CN1960&70s-CN2000&10s comparisons were included. The trait name and corresponding total GWAS loci number ($p < 1e\text{-}5$) are shown below the pie plot.



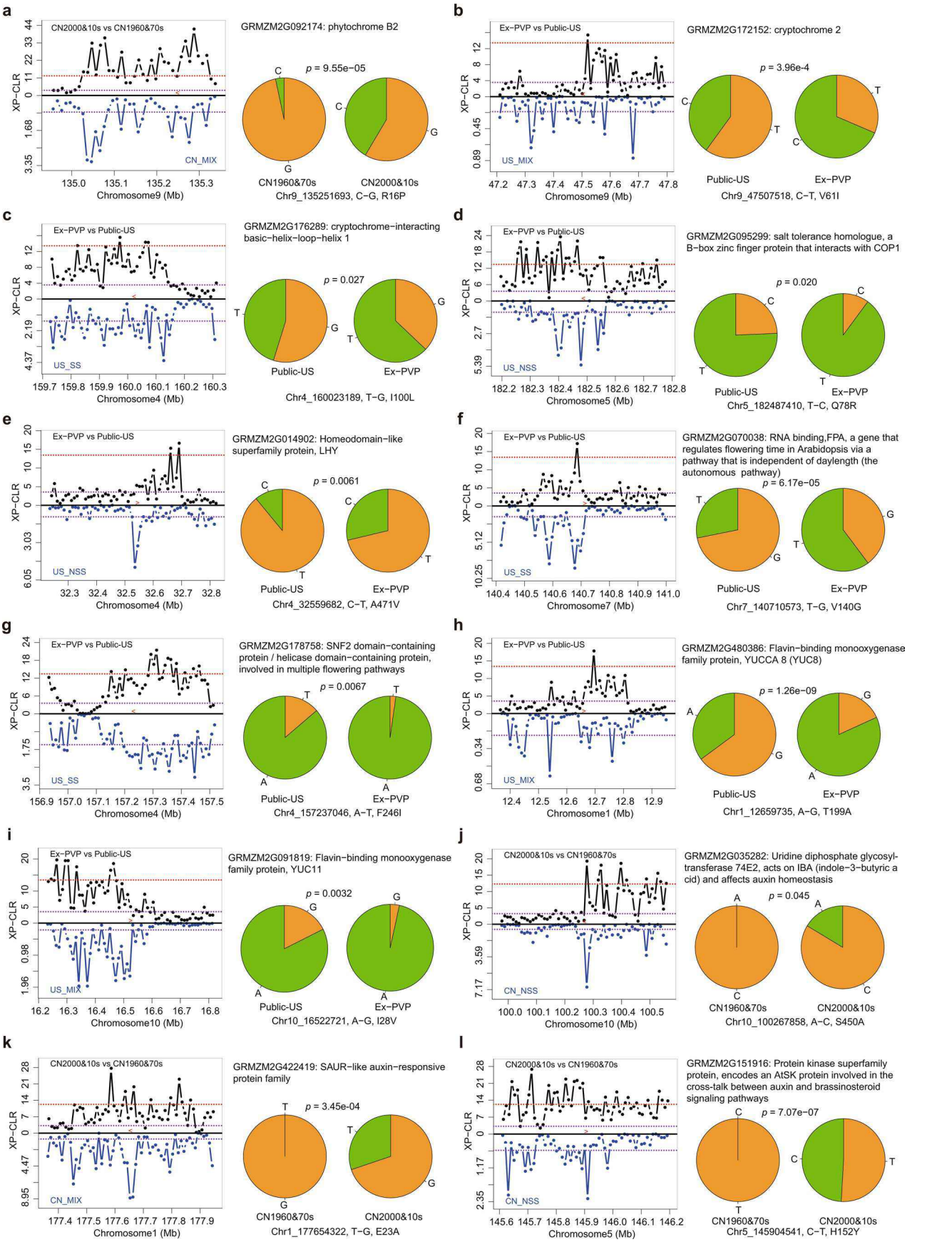
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Representative selected genes related to biotic stress responses. Each plot group represents the results for a selected representative candidate gene, which includes XP-CLR plot (left), gene annotation (above the pie plot), nonsynonymous SNP frequency changes during the corresponding breeding process (pie plot) and nonsynonymous SNP information (below the pie plot). For XP-CLR plots, the XP-CLR scores for whole data panel and subgroups are plotted above and under the zero, respectively. Red arrows along the x-axis indicate the position of the candidate genes. The blue and red horizontal dashed lines above the zero represent the 80th quantile and genome-wide significant cutoff, respectively, for XP-CLR scores in whole data panel. The horizontal dashed lines under the zero represent the 80th quantile for XP-CLR scores in subgroups. Arabidopsis homologs were used for annotation of the candidate genes. The *p*-value of fisher's exact test for allele frequency changes are shown above the pie plot. The nonsynonymous SNP information includes SNP location, variation from alleles in B73 to others, and corresponding amino acid changes (separated by comma).



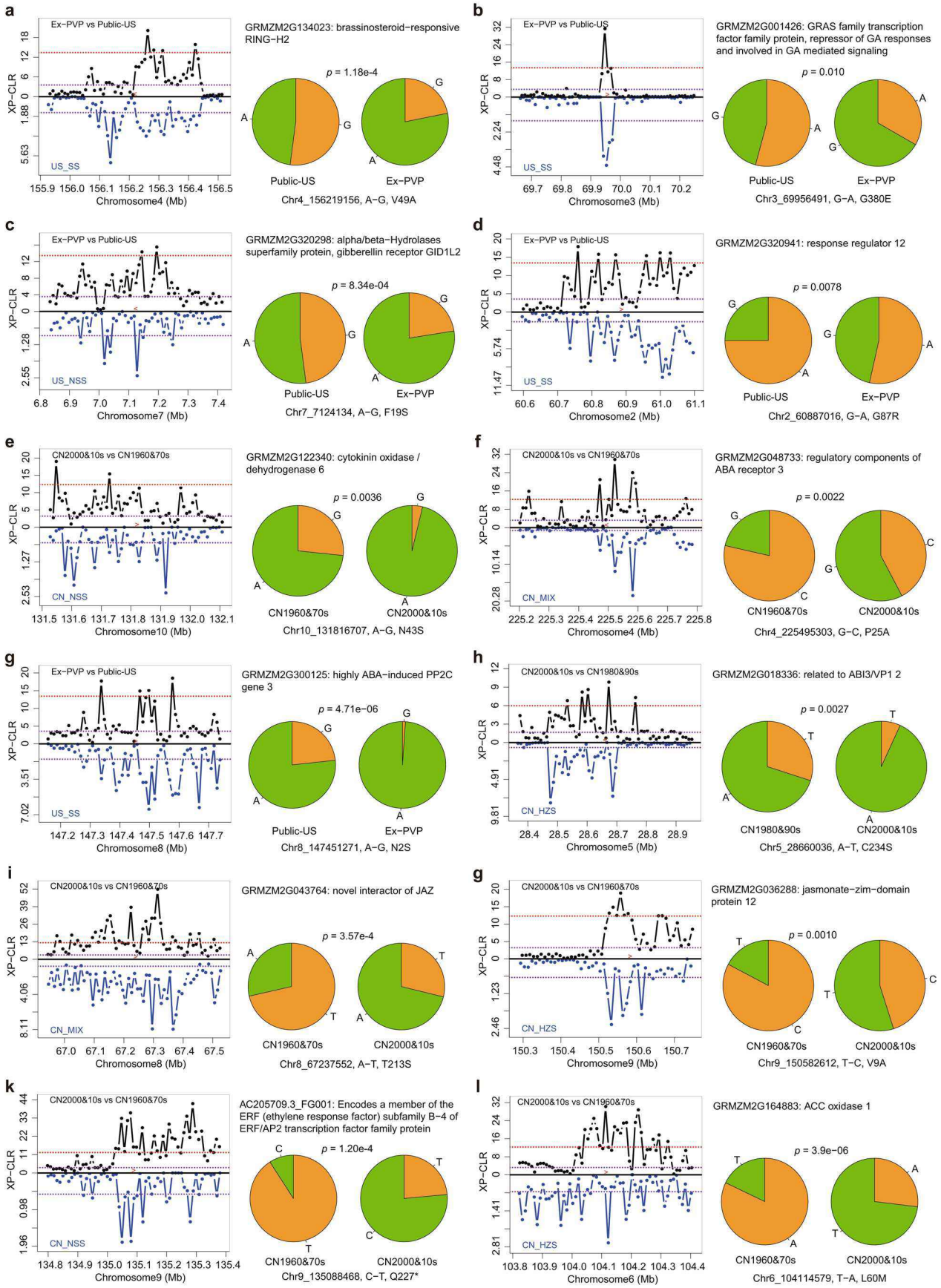
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Representative selected genes related to abiotic stress responses. Each plot group represents the results for a selected representative candidate gene, which includes XP-CLR plot (left), gene annotation (above the pie plot), nonsynonymous SNP frequency changes during the corresponding breeding process (pie plot) and nonsynonymous SNP information (below the pie plot). For XP-CLR plots, the XP-CLR scores for whole data panel and subgroups are plotted above and under the zero, respectively. Red arrows along the x-axis indicate the position of the candidate genes. The blue and red horizontal dashed lines above the zero represent the 80th quantile and genome-wide significant cutoff, respectively, for XP-CLR scores in whole data panel. The horizontal dashed lines under the zero represent the 80th quantile for XP-CLR scores in subgroups. Arabidopsis homologs were used for annotation of the candidate genes. The *p*-value of fisher's exact test for allele frequency changes are shown above the pie plot. The nonsynonymous SNP information includes SNP location, variation from alleles in B73 to others, and corresponding amino acid changes (separated by comma).



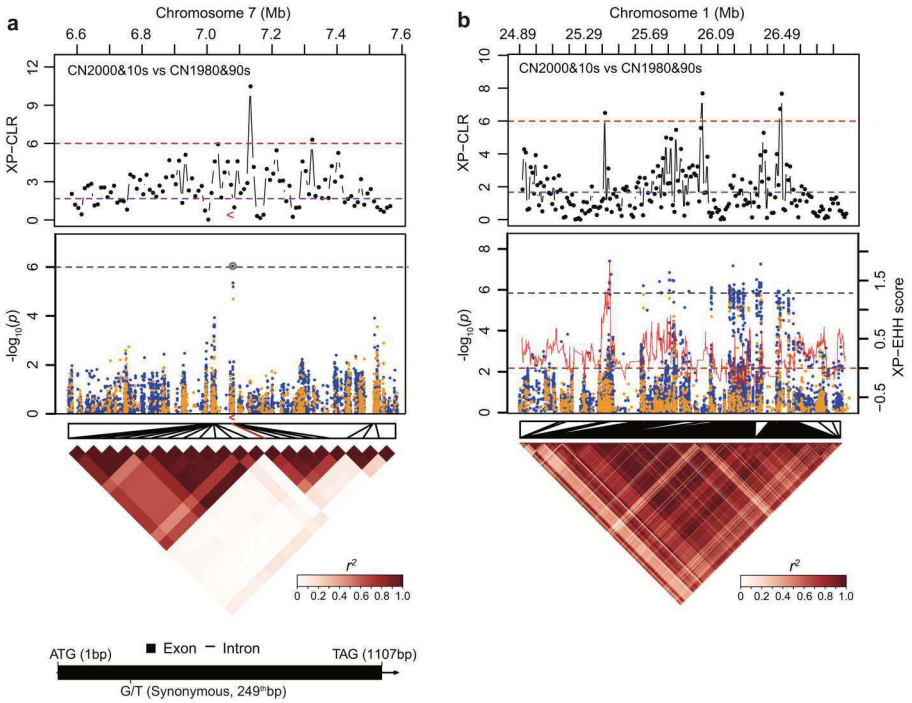
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Representative selected genes related to light signaling, flowering time regulation, biosynthesis or signaling of auxin. Each plot group represents the results for a selected representative candidate gene, which includes XP-CLR plot (left), gene annotation (above the pie plot), nonsynonymous SNP frequency changes during the corresponding breeding process (pie plot) and nonsynonymous SNP information (below the pie plot). For XP-CLR plots, the XP-CLR scores for whole data panel and subgroups are plotted above and under the zero, respectively. Red arrows along the x-axis indicate the position of the candidate genes. The blue and red horizontal dashed lines above the zero represent the 80th quantile and genome-wide significant cutoff, respectively, for XP-CLR scores in whole data panel. The horizontal dashed lines under the zero represent the 80th quantile for XP-CLR scores in subgroups. Arabidopsis homologs were used for annotation of the candidate genes. The *p*-value of fisher's exact test for allele frequency changes are shown above the pie plot. The nonsynonymous SNP information includes SNP location, variation from alleles in B73 to others and corresponding amino acid changes (separated by comma).

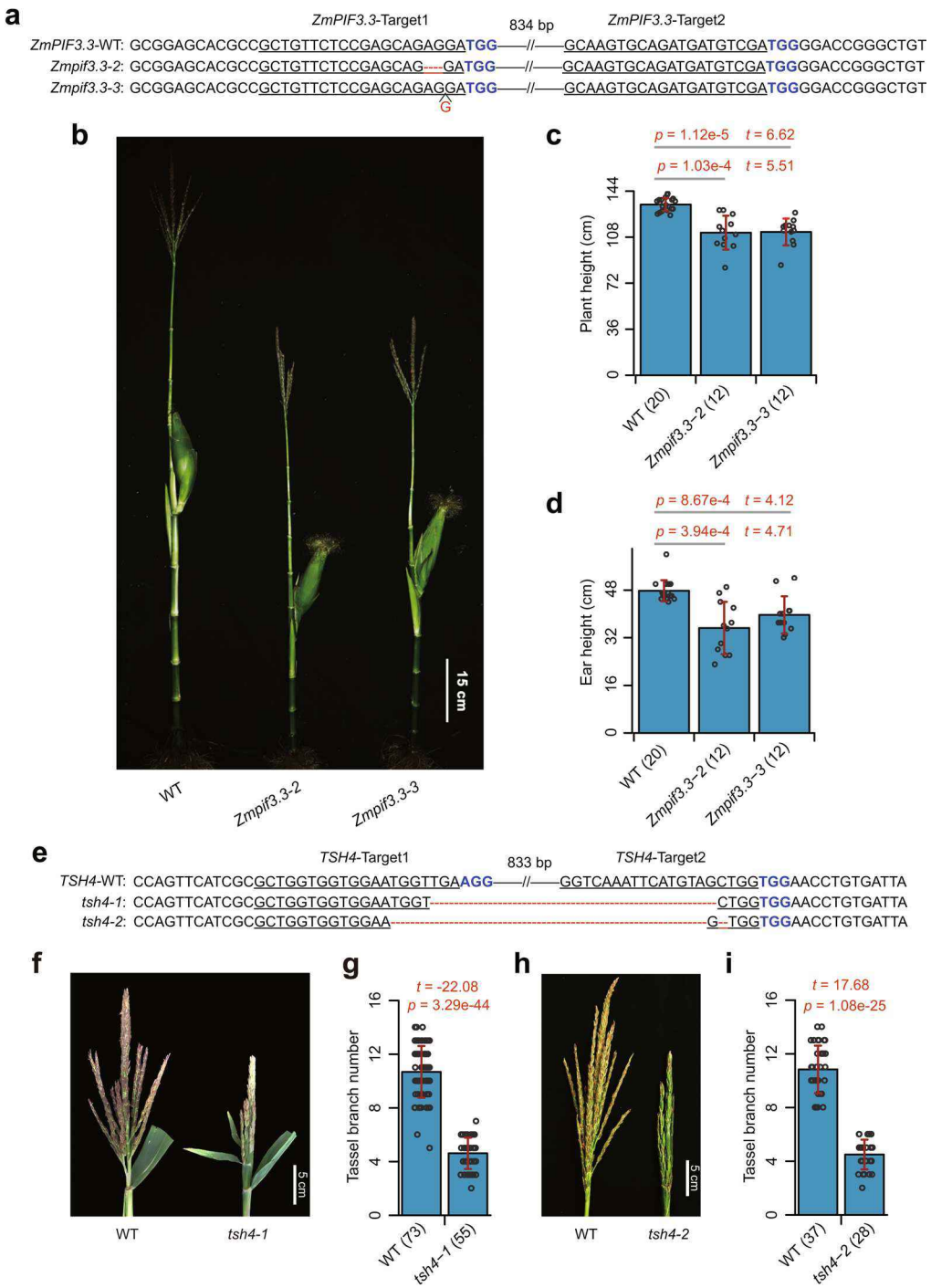


Extended Data Fig. 7 | See next page for caption.

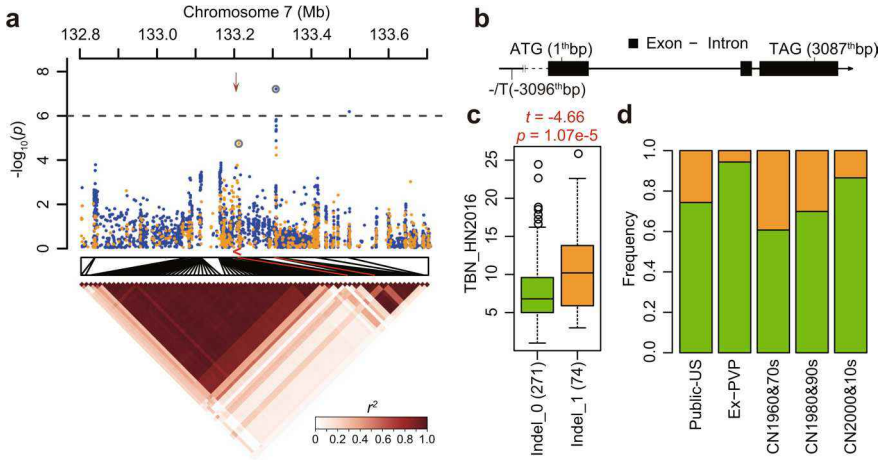
Extended Data Fig. 7 | Representative selected genes related to biosynthesis or signaling of other phytohormones. Each plot group represents the results for a selected representative candidate gene, which includes XP-CLR plot (left), gene annotation (above the pie plot), nonsynonymous SNP frequency changes during the corresponding breeding process (pie plot) and nonsynonymous SNP information (below the pie plot). For XP-CLR plots, the XP-CLR scores for whole data panel and subgroups are plotted above and under the zero, respectively. Red arrows along the x-axis indicate the position of the candidate genes. The blue and red horizontal dashed lines above the zero represent the 80th quantile and genome-wide significant cutoff, respectively, for XP-CLR scores in whole data panel. The horizontal dashed lines under the zero represent the 80th quantile for XP-CLR scores in subgroups. Arabidopsis homologs were used for annotation of the candidate genes. The *p*-value of fisher's exact test for allele frequency changes are shown above the pie plot. The nonsynonymous SNP information includes SNP location, variation from alleles in B73 to others and corresponding amino acid changes (separated by comma).



Extended Data Fig. 8 | Two detected GWAS loci for relative ear height (EP). a, **b**, XP-CLR (upper), Manhattan plot (middle) and LD heat map (lower) for the detected EP loci on 7.07 Mb of chromosome 7 (a), and 25.43 Mb of chromosome 1 (b). The candidate genes GRMZM2G398996 (a) is marked with red arrows. The structure and top SNP information of the candidate gene are shown below the LD heat map plots. To verify that the selection region on chromosome 1 might be resulted from the extended haplotype of the locus, the XP-EHH score was also investigated and shown as red curve in the Manhattan plot.



Extended Data Fig. 9 | Phenotype analyses of CRISPR/Cas9 mutations for *ZmPIF3.3* and *TSH4*. a, Sequences of *ZmPIF3.3* target regions in wild type, *Zmpif3.3-2* and *Zmpif3.3-3* CRISPR/Cas9 knockout mutants. The target sites and protospacer-adjacent motifs (PAM) are shown as underscored letters and blue letters respectively. The gap lengths of sequences are shown above the wild type sequences. **b**, Height profile of wild type, *Zmpif3.3-2* and *Zmpif3.3-3* mutant plants. Bar, 15 cm. **c**, **d**, Statistics of plant height (c) and ear height (d) of wild type, *Zmpif3.3-2* and *Zmpif3.3-3* mutant plants. **e**, Sequences of *TSH4* target regions in wild type and *tsh4* CRISPR/Cas9 knockout mutants. **f**, **g**, Tassel profile (f) and TBN statistics (g) of wild type and *tsh4-1* CRISPR-knockout mutants. Bar, 5 cm. **h**, **i**, Tassel profile (h) and TBN statistics (i) of wild type and *tsh4-2* CRISPR-knockout mutants. Bar, 5 cm. The *p*-values of two-tailed *t*-tests are shown above the plots. Error bars indicate ±s.d.



Extended Data Fig. 10 | GWAS identification of TSH4 as a candidate gene for tassel branch number (TBN) variation. a, Manhattan plot (upper left) and LD heat map (lower left) for GWAS signal TBN_7_133305039. SNP and indel based association analysis results are shown as blue and orange dots in the Manhattan plot, respectively. Peak markers and putative causal polymorphisms are circled and their positions in LD heat map are indicated by red lines. The candidate gene position in Manhattan plot is showed as red arrows. The significantly associated SNP (chr7_133305039, $P = 6.83 \times 10^{-8}$) and indel (chr7_133209283_C/CT, 1-bp deletion, $P = 1.86 \times 10^{-5}$) were strongly correlated ($r^2 = 0.53$). **b,** Candidate gene structure and polymorphisms of chr7_133209283_C/CT. **c**, **d,** Phenotype of different haplotypes (c, box plot) and haplotype frequency changes during breeding (d, bar plot), for the association signal of chr7_133209283_C/CT.

nature research

Corresponding author(s): Haiyang Wang

Last updated by author(s): Mar 5, 2020

Reporting Summary

Nature Research 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 Research policies, see Authors & Referees 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.
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/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

DNA-sequencing reads for all maize lines were deposited in the NCBI with the accession code of PRJNA609577 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA609577), and BIGD (BIG Data Center in Beijing institute of Genomics) with the accession code of CRA002372 (https://bigd.big.ac.cn/gsa/browse/CRA002372). All phenotype data of 350 inbred maize lines are included in Supplementary Table 1. All other reasonable requests for data and research materials will be made available via contacting the corresponding authors.

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 350 elite maize inbred lines (ILs), which comprise of 163 U.S. and 187 Chinese ILs, were collected for this study.
Data exclusions	During the XP-CLR analysis, to minimize the effect of population structure, we excluded the IDT germplasm (25 lines) in the selective sweep analysis as it was only utilized in the most recent breeding eras of the US and China. This was clearly described in the manuscript.
Replication	The 15 agronomic traits for 350 inbred lines were repeatedly measured across four environments. Three to four sampling replicates were used for expression analysis of ZmNAC16 and ZmSBP18, with each replicate consists of leaf collar tissues from 3 independent plants.
Randomization	A randomized complete block design was used in all four trials for phenotype collection.
Blinding	The investigators were blinded to the maize lines during data collection.

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
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

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