Genome-wide selection and genetic improvement during modern maize breeding

Baobao Wang1,2, Zechuan Lin2,3, Xin Li1,3, Yongping Zhao1, Binbin Zhao1, Guangxia Wu1, Xiaojing Ma1, Hai Wang1, Yurong Xie1, Quanquan Li1,3, Guangshu Song1, Dexin Kong2, Zhigang Zheng5, Hongbin Wei3, Rongxin Shen3, Hong Wu3, Cuixia Chen4, Zhaodong Meng5, Tianyu Wang7, Yu Li7, Xinhai Li7, Yanhui Chen8, Jinheng Lai8,9, Matthew B. Hufford10, Jeffrey Ross-Ibarra11, Hang He2,12,13 and Haiyang Wang1,5,12,13

Since the development of single-hybrid maize breeding programs in the first half of the twentieth century, maize yields have increased over sevenfold, and much of that increase can be attributed to tolerance of increased planting density. 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. Previous studies have investigated the genome-wide changes that occurred during maize domestication and expansion to novel environments but analysis of genetic improvement during modern breeding has so far been limited in scope. 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, 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. 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, 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...
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 (\(RVE1\))\(^{24,27}\), 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...
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 (Supplementary Table 9) and is consistent with models of selection on polygenic traits\(^{10}\). 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\(^{11}\). We thus used the cross-population composite likelihood ratio approach (XP-CLR)\(^{12}\) 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)\(^{8}\) but comparable to sweeps for improvement (average size 176 kb)\(^{8}\) and tropical–temperate adaptation (average size 150.9 kb)\(^{8}\). 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.
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<sup>46</sup>. 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<sup>43,45</sup>. ZmPHYB2 (GRMZM2G092174) and two phytochrome-interacting factors, ZmPIF4 (GRMZM5G865967) and ZmPIF3.3 (GRMZM2G062541; ref. <sup>48</sup>) 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.
critical roles in regulating various morphological traits in maize. Notably, our GWAS identified a SNP (chr7_133305039, \( P = 6.83 \times 10^{-4} \)) 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 TSH4 promoter. Association signals for SNP and indel are shown as blue dots and dashed lines represent the genome-wide cutoff, 80th quantile for XP-CLR score, and median of genome-wide distribution. The SNP-1245 is shown in red.

**Fig. 5 | Validation of two candidate genes associated with EH and TBN.** a, XP-CLR (above x axis) and \( \pi \)-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 distribution. The SNP-1245 is shown in red. 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 promoter fragment but not that of TSH4-DNA. l, Luciferase activity assay shows that coexpression of RA2 effectively inhibits the LUC reporter gene driven by the TSH4 promoter but not the TSH4 promoter. m, Box plot for TBN of different haplotypes of the ZmPIF3.3-1 Target1 (94), Target2 (55), Hap2 (94), Hap3 (16), and Hap4 (13). 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 (\( P = 6.31 \times 10^{-4} \); 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. Notably, our GWAS identified a SNP (chr7_133305039, \( P = 6.83 \times 10^{-4} \)) 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, 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 1245bp 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 factors89. 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 meristems89. Yeast one-hybrid assay showed that RA2 could directly bind to the TSH4 promoter fragment containing wild-type CGGC motif (B73 type, TSH4β3) but not the TSH4 promoter fragment containing mutated CGGC motif (Chang-7 type, TSH4α2) (Fig. 5j,k). Luciferase activity assay showed that coexpression of RA2 effectively inhibited the LUC reporter gene driven by the TSH4β3 promoter but not the TSH4α2 promoter (in which the CGGC motif was mutated to the Chang-7 type, Fig. 5i). These results are consistent with the observations that RA2 and TSH4 have complementary expression domains in the main inflorescence89. 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; detablined 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

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
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 on this locus). The leaf collars of expanded and unexpanded leaves from V7 seedlings were collected for DNA 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 lm4 package to obtain a best linear unbiased predictor (BLUP) value as follows:

\[ y_i = \mu + \beta_i + \epsilon_i + \psi_i \]

where \( \mu \) is the mean, \( \beta_i \) is the genotype effect of the \( i \)-th inbred, \( \epsilon_i \) is the effect of the \( j \)-th environment, \( \psi_i \) is the genotype–environment interaction and \( \epsilon_{ij} \) is the error term. The analysis was performed for each environment, and the genotype–environment interactions were used to obtain a best linear unbiased predictor of the traits for each inbred line.

DNA isolation and sequencing. Six inbred lines (Mo17, Zheng58, Chang7-2, 478, 5003 and 8112) have been previously sequenced to more than 20x coverage (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, producing 150-bp paired-end reads (~10.4 terabases, depth ~99.5%). We selected 11,622,737 SNPs (MAF > 0.05) from 350 inbred lines and used these SNPs to perform GWAS of all traits. The missing genotypes \( \theta = 0.05 \) 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.fst.gz) generated on the basis of the B73 reference genome (GCA_000050055.5_B73_RefGen_v3) by using SnapEff (v.4.3a). We found that 7,469,668 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 0.690 reported in maize (https://www.ncbi.nlm.gov). This is similar to an earlier report using an elite inbred lines panel 16.

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

LD (calculated as \( r^2 \)) was also calculated with PLINK 1.90b3.42) with the following parameters:

- \( \text{LD-window} = 20 \times \text{ld-window} \text{99999} \times \text{ld-window} \text{kb} \text{ 500} \),
- \( \text{LD-window} \times \text{2} = \text{ld-window} \text{kb} \text{ 500} \),
- the distance of \( \text{LD} \) decay was represented as the physical distance over which \( r^2 \) drops to 0.2.

We first conducted principle component analysis (PCA) with Eigensoft using all 25,320,665 SNPs and selected the first three principle components (PCs), explained 19.16% of the genome-wide variation (Supplementary Fig. 2). We further performed a similar principle component analysis for the complete set of 10-kb windows and selected the first two principle components (PC2), explaining 55.83% of the genome-wide variation.

We then used Principal Component Analysis (PCA) to calculate the first two principle components (PCs). The first two principle components (PCs) were used for further analysis to evaluate the nucleotide diversity of inbred lines of every era.

Select mean values for all traits were calculated for each environment. For example, the mean value for the trait Fst in environment 1 was calculated as the average of the trait values for all inbred lines in environment 1. The mean value for the trait Fst in environment 2 was calculated as the average of the trait values for all inbred lines in environment 2. This process was repeated for all environments.

GWAS of plant morphological traits. We selected 11,622,737 SNPs (MAF > 0.05 and missing rate <30%) to perform GWAS of all traits. The missing genotypes were imputed using the method implemented in the ADMIXTURE tool 16. We used 280,000 SNPs for the population genetic analyses (PCA) and 25,320,665 SNPs for the GWAS of all traits. The reported values were calculated using the following formula:

\[ r^2 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \]
were imputed with Beagle (v4.1) with default parameters. The GWAS was conducted with a linear mixed model that was implemented in the EMMAX package. We performed GWAS using both the BLUP and single trait 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 r² > 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 × 10⁵ (Benjamini–Hochberg FDR < 0.05) as the genomewide significant cutoff. We determined significant signals with the following two criteria: (1) the P values of the signals for BLUP values were < 1 × 10⁻⁵ or (2) the P values of signals were consistently lower than 1 × 10⁻⁵ for at least two environmental traits. For adjacent GWAS loci (<500 kb), loci independence was determined by pairwise linkage analysis of significant SNPs (if r² < 0.5, they were declared independent). The confidence intervals of the GWAS loci were determined by local LD block analysis where pairwise r² of the SNPs with P < 1 × 10⁻⁵ should be > 0.3. Genes located directly in or within 33 kb (genome-wide average distance of LD decay to r² = 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 the MLM package. 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 19). Since the positions of 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 CN1980890s, Public-Us to CN2000110s, Ex-Pv to CN1980890s and Ex-Pv to CN2000110s) using the four taxa approach (Extended Data Fig. 9).

Phenotypic 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 introgression, we analyzed the extent of genetic introgression from the US inbred lines to the Chinese inbred lines (Public-Us to CN1980890s, Public-Us to CN2000110s, Ex-Pv to CN1980890s and Ex-Pv to CN2000110s) using the four taxa approach which calculates the extremely shared derived variants between two taxa (ABRA–BABA statistic, also known as f₂ statistic). We selected 65 tropical and subtropical maize inbred lines from the maize HapMap 3 (ref. 19) as the outgroup, the inbred lines from the breeding era of CN1960870s as control and screened the potential introgressed regions with a window size of 100 kb. Windows with meaningless result (f₂ > 1, f₂ < 0 or with Patterson’s D statistic < 0) were removed and these with strongest 5% of f₂ 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 PRJNA695775 and BIGD (Big Data Center in Beijing Institute of Genomics) with the accession code of CRA052702. 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 relevant requests for data and research materials are available via contacting the corresponding authors.

References

**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**

**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.
Reprints and permissions information is available at www.nature.com/reprints.
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 $p < 1 \times 10^{-6}$ and $1 \times 10^{-5}$, 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 ($p < 1 \times 10^{-5}$) are shown below the pie plot.
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 | 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**, 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**, Tassel profile (f) and TBN statistics (g) of wild type and tsh4-1 CRISPR-knockout mutants. Bar, 5 cm. 

**h**, 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.
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

<table>
<thead>
<tr>
<th>Data collection</th>
<th>We used open source software and codes for data collection.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data analysis</td>
<td>All softwares used in the present study are publicly available and the corresponding versions are described in detail in the Online Methods.</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>A total of 350 elite maize inbred lines (ILs), which comprise of 163 U.S. and 187 Chinese ILs, were collected for this study.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data exclusions</td>
<td>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.</td>
</tr>
<tr>
<td>Replication</td>
<td>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.</td>
</tr>
<tr>
<td>Randomization</td>
<td>A randomized complete block design was used in all four trials for phenotype collection.</td>
</tr>
<tr>
<td>Blinding</td>
<td>The investigators were blinded to the maize lines during data collection.</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Materials &amp; experimental systems</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>☒ Involved in the study</td>
<td>Involved in the study</td>
</tr>
<tr>
<td>☐ Antibodies</td>
<td>☒ ChIP-seq</td>
</tr>
<tr>
<td>☒ Eukaryotic cell lines</td>
<td>☒ Flow cytometry</td>
</tr>
<tr>
<td>☒ Palaeontology</td>
<td>☒ MRI-based neuroimaging</td>
</tr>
<tr>
<td>☒ Animals and other organisms</td>
<td></td>
</tr>
<tr>
<td>☒ Human research participants</td>
<td></td>
</tr>
<tr>
<td>☒ Clinical data</td>
<td></td>
</tr>
</tbody>
</table>