nature genetics

Genome-wide genetic changes during modern breeding of maize

Yinping Jiao[1,2], Hainan Zhao[1,2], Longhui Ren[1,2], Weibin Song[1,2], Biao Zeng[1], Jinjie Guo[1], Baobao Wang[1], Zhipeng Liu[1], Jing Chen[1], Wei Li[1], Mei Zhang[1], Shaojun Xie[1] & Jinsheng Lai[1]

The success of modern maize breeding has been demonstrated by remarkable increases in productivity over the last four decades. However, the underlying genetic changes correlated with these gains remain largely unknown. We report here the sequencing of 278 temperate maize inbred lines from different stages of breeding history, including deep resequencing of 4 lines with known pedigree information. The results show that modern breeding has introduced highly dynamic genetic changes into the maize genome. Artificial selection has affected thousands of targets, including genes and non-genic regions, leading to a reduction in nucleotide diversity and an increase in the proportion of rare alleles. Genetic changes during breeding happen rapidly, with extensive variation (SNPs, indels and copy-number variants (CNVs)) occurring, even within identity-by-descent regions. Our genome-wide assessment of genetic changes during modern maize breeding provides new strategies as well as practical targets for future crop breeding and biotechnology.

Maize is one of the most important crops in the world. After domestication from teosinte (*Zea mays* ssp. *Parviglumis*) around 10,000 years ago[1] and a long period afterward of breeding by farmers, maize has undergone extensive scientific breeding in recent years. Modern breeding efforts over the last few decades have led to a remarkable yield increase for this crop[2,3]. Maize is exceptionally diverse[4], and the pattern of genome-wide genetic variation among a number of maize lines has recently been reported[5,6]. To assess genetic changes during breeding over the last few decades, we sequenced the whole genomes of 278 lines, including 90 Ex-PVP lines (lines with expired Plant Variety Protection Act certificates), 36 public US lines (publically avaliable, non-PVP lines) and 152 elite Chinese lines (**Supplementary Table 1**). These lines represent an extensive collection of the most advanced publically available temperate maize inbred lines.

A total of 1.3 trillion base pairs of data comprising 13 billion 100-bp reads was generated, with an average sequencing depth of ~2× for each line. In analyzing the data, 27,818,705 SNPs were identified. A subset of 6,686,326 SNPs with a missing data rate of less than 50% in the population was used for subsequent analysis (**Supplementary Table 2**). A total of 1,015,790 SNPs were found in genes, and 283,186 SNPs were found in coding sequences. We detected 158,296 nonsynonymous and 138,918 synonymous SNPs in coding regions; the nonsynonymous-to-synonymous ratio was 1.14. We identified 3,046 large-effect SNPs (including SNPs in start codons, stop codons and exon-intron splice sites) in 2,282 genes. To validate SNP quality, we compared data from three lines (Hp301, Mo17 and P39) to sequences in maize HapMap1 (ref. 5), finding over 95.6% accordance (out of 207,825 overlapping sites). SNPs were further verified through a genome-wide association study (GWAS) for three traits (cob color, silk color and date to anthesis) after SNP imputation. The top signals of the GWAS for these three traits included the expected targets known to influence these traits. Examples of this included the identification of a SNP in the tandem repeat region of *p1* (ref. 7) on chromosome 1 for cob color (a similar result was obtained in a previous GWAS for this trait)[8], a SNP located 596 bp away from *r1* (ref. 9) on chromosome 10 for silk color and a SNP 1.2 Mb away from *Vgt1* (ref. 10) on chromosome 8 for date to anthesis (**Fig. 1**, **Supplementary Fig. 1** and **Supplementary Table 3**).

The overall nucleotide diversities (π) of these 278 lines were lower than previously reported in a more diverse population (0.006)[5]. Comparison of the Ex-PVP group and the public US group, two populations that are both representative of temperate maize (**Fig. 2**) but are separated by approximately 25 years of breeding history, clearly indicated that the Ex-PVP group had 26% less nucleotide diversity (π = 0.0039) than their ancestral US lines (π = 0.0053). The elite Chinese lines and public US lines had nearly the same level of nucleotide diversity (**Supplementary Table 4**) and showed little genetic differentiation (F_{ST} = 0.023). All three

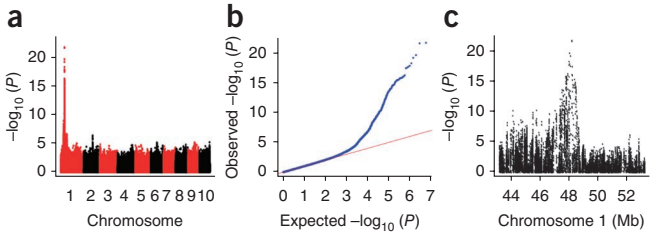

Figure 1 GWAS results for cob color. (**a**) Manhattan plot. (**b**) Quantile-quantile plot. (**c**). Regional Manhattan plot of 5 Mb on either side of the peak SNP.

[1]State Key Laboratory of Agrobiotechnology and National Maize Improvement Center, China Agricultural University, Beijing, People's Republic of China.
[2]These authors contributed equally to this work. Correspondence should be addressed to J.L. (jlai@cau.edu.cn).

Received 13 December 2011; accepted 7 May 2012; published online 3 June 2012; doi:10.1038/ng.2312

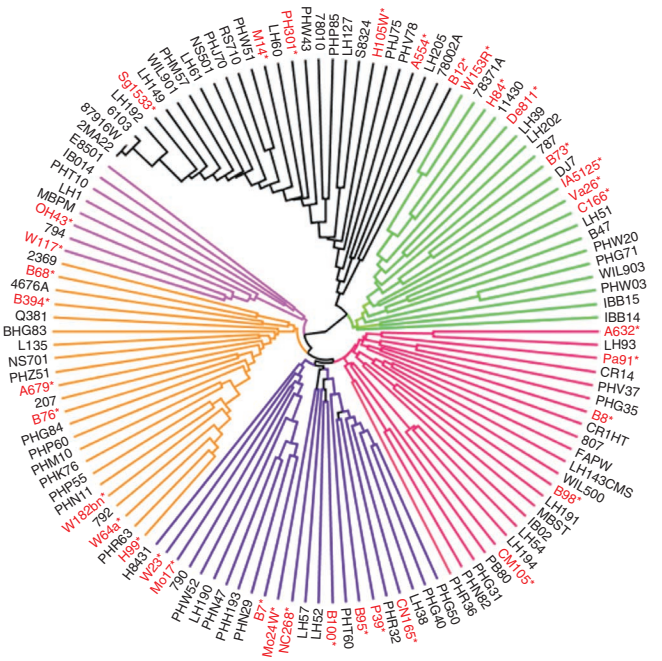

Figure 2 Neighbor-joining tree of the 126 US maize inbred lines. Lines in the public US group are shown in red followed by an asterisk. Ex-PVP lines are shown in black.

groups contained excess amounts of rare alleles, a finding supported by negative Tajima's *D* values (**Supplementary Table 4**), suggesting an ongoing expansion of the population during modern maize breeding. The lower Tajima's *D* value of the entire genome in these three populations compared to genic regions (**Supplementary Table 4**) suggests that stronger selection has occurred in non-genic regions.

In order to identify genomic targets of artificial selection, we screened for signals of selective sweeps in the three groups separately using a composite likelihood ratio (CLR) approach[11]. Using a threshold by which the top 1% of CLR values are selected, we identified 405 regions in the public US group, 407 in the Ex-PVP group and 408 in the elite Chinese group as candidate regions that have experienced a selective sweep. These regions accounted for 2.38%, 2.10% and 1.74% of the maize genome, respectively (**Fig. 3**, **Supplementary Fig. 2** and **Supplementary Table 5**). These target regions had lower levels of nucleotide diversity and extremely negative Tajima's *D* values (**Supplementary Table 4**). Although most targets mapped to protein-coding regions, a number of target regions did not (54 targets for the public US group, 48 for the Ex-PVP group and 65 for the elite Chinese group), which suggests an effect of artificial selection on non-genic regulatory elements. These targets showed little overlap with previously identified domestication or improvement loci[12,13], indicating that most of these targets may have emerged from more advanced stages of maize breeding. There were a total of 1,835 genes from the maize filtered-gene model within the target regions (529 in the elite Chinese group, 689 in the public US group and 763 in the Ex-PVP group; **Supplementary Table 5**). The functions of some of these genes have previously been reported in maize (**Supplementary Table 6**).

Notably, only a small proportion of the selection targets identified in the Ex-PVP and the public US groups overlapped. A total of 51 targets identified in the US group (12.6%) fell within the 1% tails of CLR in the Ex-PVP group, and 149 (36.8%) fell within the 5% tails. Similarly, 128 targets identified in the Ex-PVP group (31.4%) fell within the 5% tails of CLR in the US group. Our data suggest that different stages of maize improvement could have targeted different genomic regions, with targets in the early stages of breeding potentially fixed in the population of later stages of breeding. Similarly, a limited number of targets were shared between the Chinese and US maize lines. A total of 207 targets of selection in the Chinese group (50.7%) fell within the 5% tails of CLR of the US group and/or the Ex-PVP group, and 116 targets (28.4%) in the US group and 137 targets (33.8%) in the Ex-PVP group fell within the 5% tails of CLR of the Chinese group. Limited sharing of candidate regions of selection between populations from different geographic regions is also observed in humans[14]. The lack of shared regions might be because the Chinese and US maize lines underwent different selection pressures to adapt to local agricultural conditions. An alternative explanation is that the same agronomic trait may be obtained by selection on different genomic regions. For example, seed size in rice is known to be controlled by multiple genes[15–19].

Genetic changes within a breeding program were explored by deeply sequencing the genomes of four inbred lines (5003, 8112, 478 and Zheng58) with known breeding pedigree information (**Supplementary Fig. 3**). These four lines, which are sampled from three generations of breeding, represent two major breeding advances that gave rise to the female parent of the most widely planted hybrid in China. A total of 256 Gb of data was obtained for the four lines, with an average genome coverage of 27× (**Supplementary Table 7**), including previous data that reported 5× coverage of a subset of these lines[6].

Mapping the reads of these four deep-sequenced lines to the maize B73 reference genome[20] identified a total of 5,058,396 SNPs that were covered by at least 5 reads, with 895,527 located in genic regions (31,262 genes). Among the 216,923 SNPs in coding regions, 2,153 were found in the large-effect sites of 1,986 genes (**Supplementary Table 8**). Additionally, 287,504 short indels of 1–10 bp in length were identified in the 4 inbred lines. A total of 2,595 of these indels were located in protein-coding regions (**Supplementary Table 9**). Most

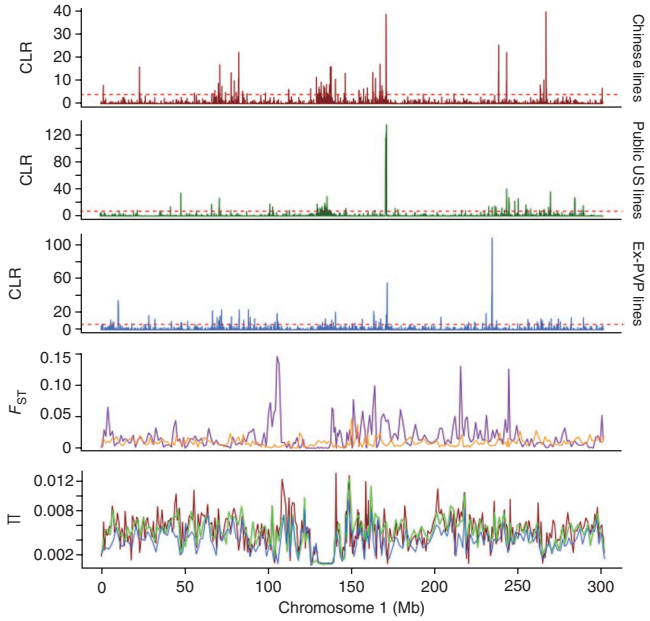

Figure 3 CLR and genetic diversity of chromosome 1 in public US, Ex-PVP and elite Chinese maize groups. Green lines, public US group; blue lines, Ex-PVP group; red lines, elite Chinese group; orange lines, F_{ST} of the Chinese and US (both public and Ex-PVP) maize groups; purple lines, F_{ST} of the public US group and Ex-PVP groups.

Table 1 Mutation rates of base substitutions, short indels and copy-number variations

	5003→478[a]	8112→478[a]	478→Z58[b]
Number of years since breeding	26	26	24
Number of sites surveyed (all three generation with sequencing depth of ≥5)	176,528,421	341,563,217	518,091,638
Number of single-base mutations	658	663	1,312
Number of single-base mutations in genic regions	197	206	133
Number of single-base mutations in coding regions	80	81	32
Number of nonsynonymous single-base mutations	41	37	21
Single-base substitution rate (per site per year)	7.17×10^{-8}	3.73×10^{-8}	5.28×10^{-8}
Number of 1–10 bp indels	708	825	1,609
Mutation rate of indels of 1–10 bp (per site per year)	7.71×10^{-8}	4.64×10^{-8}	6.47×10^{-8}
Number of genes surveyed	5,256	7,860	13,116
Number of CNV genes	225	292	590
CNV mutation rate (per gene per year)	8.23×10^{-4}	7.14×10^{-4}	9.37×10^{-4}

[a]Only the mutations transferred from 478 to Z58 were considered. [b]Only the mutation sites showing no polymorphism in comparison of 8112 and 478 or 5003 and 478 were considered.

of the indels identified were 1 bp in length (**Supplementary Fig. 4**). The SNPs called from deep-resequencing data were validated through comparison with a local *de novo* assembly of genic regions (99.42% concordance; **Supplementary Table 10**). Validation with simulative reads generated from nine Mo17 BAC sequences showed an accuracy of 99.1% for SNPs and 95.0% for indels. Comparison with the nine Mo17 BAC sequences also suggested that our pipeline had missed 36.9% of SNPs and 76.1% of indels. Most of the missed SNPs and indels were located in SNP or indel clusters (for example, regions with more than three SNPs within 10 bp).

Identity-by-descent (IBD) regions inherited across breeding generations can be used to estimate the rate of genetic changes during the breeding process (**Supplementary Fig. 3**). There were IBD blocks originating from 5003 that were inherited throughout all three generations. Within these blocks, regions of a total of 176 Mb covered by at least five reads in all three generations were found (**Table 1**). In these regions, 658 substitutions were identified between 5003 and 478, and all remained unchanged from 478 to Zheng58, leading to an estimated nucleotide substitution rate of 7.17×10^{-8} per site per year. Similarly, regions of 341 Mb originally from 8112 were identified. A total of 663 substitutions between 8112 and 478 in these regions gave an estimated substitution rate of 3.73×10^{-8} per site per year. To rule out the possibility of pre-existing heterozygosity in the ancestral 5003 or 8112 lines, we used regions totaling 518 Mb in length in which there was no polymorphism between 5003 and 478 or between 8112 and 478 but for which there were polymorphisms between 478 and Zheng58. A total of 1,312 SNPs that were different between 478 and Zheng58 in these 518-Mb regions gave an estimate of 4.91×10^{-8} substitutions per site per year. The average rate from the three estimates is 5.39×10^{-8} substitutions per site per year. In genic regions, the average rate was 4.79×10^{-8} substitutions per site per year. Intergenic regions, being rich in transposons and repetitive sequences, are hypermethylated[21] and could have higher mutation rates. The estimate of whole genome average single-base mutation rate was slightly higher than previous estimates for *tb1* between maize and teosinte[22] and in humans[23], and is ten times more than estimates in *Arabidopsis thaliana*[24] and *Caenorhabditis elegans*[25]. Nevertheless, the transition/transversion ratio (2.5) in maize was very similar to that observed in *Arabidopsis* (2.4)[24], with the highest mutation rate found for GC>AT transversions[26] (**Supplementary Fig. 5**).

Similarly, there were 3,142 indels of 1–10 bp in length found in the IBD regions of the three generations. A total of 716 of these indels, newly generated during breeding, were located in genic regions (86% were 1 or 2 bp in length, causing changes to the encoded protein sequences). The average mutation rate of short indels was estimated to be 6.13×10^{-8} substitutions per site per year (**Table 1**), which is higher than that described in *Arabidopsis*[24].

We found that 8.5% of maize genes from the filtered-gene model (3,305 genes) had CNVs among the four genomes. The average CNV rate calculated by us (8.57×10^{-4} per gene per year) was lower than that described in humans (1.2×10^{-2})[27]. We note that our CNV analysis was focused on genic regions, which may partially explain these differences. Compared to its two parents, 5003 and 8112, the inbred 478 line showed altered copy numbers in most CNV-containing genes. There were 333 genes in inbred 478 with copy numbers that were higher or lower than in either parent (**Supplementary Table 11**).

The high frequency of *de novo* genetic changes identified in IBD regions suggests that many new alleles were generated during the breeding process. As suggested from a study in soybean[28], these newly acquired alleles can potentially have phenotypic implications.

Rare alleles, like those reported in the *DGAT* gene of maize[29] and the *NAM-B1* gene of wheat[30], are very important sources of genetic improvement made through breeding. To further investigate the association of rare alleles with important agricultural traits, we inspected the allele frequencies of SNPs reported to be associated with agronomic quantitative trait loci (QTLs). Of 173 QTL-associated SNPs identified from a GWAS of the nested association mapping (NAM) population[8,31,32] and detected in our resequenced population, 63 (36.4%) had allele frequencies of less than 0.05 (**Supplementary Fig. 6**).

The availability of sequence information for 278 maize lines and a set of deep-sequenced lines allowed us to quantify changes in rare alleles during the breeding process. We found apparent accumulation of rare alleles during breeding, with 55% of segregating sites being rare in the Ex-PVP group, contrasting with 38% in the public US group. Similarly, the proportions of rare alleles in elite maize lines have continuously increased from 0.8% to 4.61% following advances in breeding (**Fig. 4**).

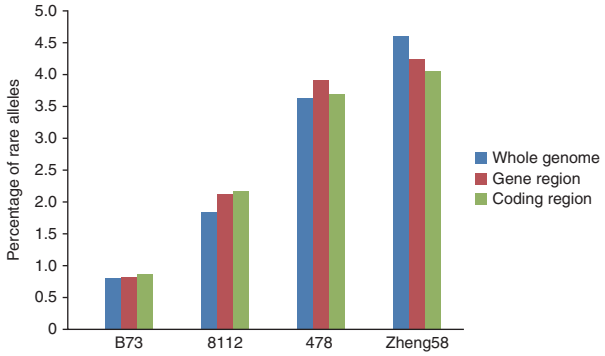

Figure 4 The percentage of rare alleles in four related inbred lines. Allele frequencies were calculated in the 278 Chinese and US lines. Rare alleles are defined as those present in ≤5% of the sequenced lines. B73 is known to be the ancestral line of 8112, 478 and Zheng58. All share the Iowa Stiff Stalk background.

Our results suggest that the relative fraction of rare alleles can potentially be used as a selection index in future breeding programs, which may reduce the time and effort required in large-scale field tests, particularly as the costs of genotyping become reasonably low. Additionally, genes identified within the breeding target regions, especially those in the Ex-PVP group, might be directly applied to future breeding or biotechnology programs. The SNP data from elite inbreds will also be useful when new breeding technologies, such as genome selection[33,34], come of age in maize.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequencing data from 278 maize inbred lines has been deposited in the NCBI Sequence Read Archive (SRA) database (SRA049859). Contigs with length of more than 200 bp generated in assembling 1,000 genes from the four deep-sequenced inbred lines have been deposited in NCBI GenBank JQ886798–JQ887980).

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank E.S. Buckler and J. Ross-Ibarra for helpful discussions, E.S. Buckler, T.R. Rocheford, M. Bohn and P. Becraft for assistance in making some of the Ex-PVP lines available and J. Dai, S. Wang and T. Wang for sharing Chinese germplasm. Research is supported by the National Basic Research Program (973 program) (2009CB118400).

AUTHOR CONTRIBUTIONS

J.L. designed the project. J.L., Y.J. and H.Z. wrote the manuscript. Y.J., H.Z., L.R., B.Z. and S.X. performed most data analyses. W.S., J.G., B.W., Z.L., J.C., W.L. and M.Z. collected the inbred lines and prepared DNA samples for sequencing.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2312.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Doebley, J. The genetics of maize evolution. *Annu. Rev. Genet.* **38**, 37–59 (2004).
2. Duvick, D.N. The contribution of breeding to yield advances in maize (*Zea mays* L.). *Adv. Agron.* **86**, 83–145 (2005).
3. Duvick, D.N. Commercial strategies for exploitation of heterosis. in *The Genetics and Exploitation of Heterosis in Crops* (eds. Coors, J.G. & Pandey, S.) 295–304 Misc: (ASA-CSSA-SSSA Publication, Madison, Wisconsin, 1999).
4. Liu, K. *et al.* Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. *Genetics* **165**, 2117–2128 (2003).
5. Gore, M.A. *et al.* A first-generation haplotype map of maize. *Science* **326**, 1115–1117 (2009).
6. Lai, J. *et al.* Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat. Genet.* **42**, 1027–1030 (2010).
7. Grotewold, E., Athma, P. & Peterson, T. A possible hot spot for Ac insertion in the maize *P* gene. *Mol. Gen. Genet.* **230**, 329–331 (1991).
8. Tian, F. *et al.* Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* **43**, 159–162 (2011).
9. Scanlon, M.J., Stinard, P.S., James, M.G., Myers, A.M. & Robertson, D.S. Genetic analysis of 63 mutations affecting maize kernel development isolated from Mutator stocks. *Genetics* **136**, 281–294 (1994).
10. 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).
11. Nielsen, R. *et al.* Genomic scans for selective sweeps using SNP data. *Genome Res.* **15**, 1566–1575 (2005).
12. 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).
13. Wright, S.I. *et al.* The effects of artificial selection on the maize genome. *Science* **308**, 1310–1314 (2005).
14. Pickrell, J.K. *et al.* Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* **19**, 826–837 (2009).
15. Song, X.J., Huang, W., Shi, M., Zhu, M.Z. & Lin, H.X. A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. *Nat. Genet.* **39**, 623–630 (2007).
16. Weng, J. *et al.* Isolation and initial characterization of *GW5*, a major QTL associated with rice grain width and weight. *Cell Res.* **18**, 1199–1209 (2008).
17. Shomura, A. *et al.* Deletion in a gene associated with grain size increased yields during rice domestication. *Nat. Genet.* **40**, 1023–1028 (2008).
18. Fan, C. *et al. GS3*, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. *Theor. Appl. Genet.* **112**, 1164–1171 (2006).
19. Wang, E. *et al.* Control of rice grain-filling and yield by a gene with a potential signature of domestication. *Nat. Genet.* **40**, 1370–1374 (2008).
20. Schnable, P.S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
21. Palmer, L.E. *et al.* Maize genome sequencing by methylation filtration. *Science* **302**, 2115–2117 (2003).
22. Clark, R.M., Tavare, S. & Doebley, J. Estimating a nucleotide substitution rate for maize from polymorphism at a major domestication locus. *Mol. Biol. Evol.* **22**, 2304–2312 (2005).
23. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
24. Ossowski, S. *et al.* The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**, 92–94 (2010).
25. Denver, D.R. *et al.* A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. *Proc. Natl. Acad. Sci. USA* **106**, 16310–16314 (2009).
26. You, Y.H., Li, C. & Pfeifer, G.P. Involvement of 5-methylcytosine in sunlight-induced mutagenesis. *J. Mol. Biol.* **293**, 493–503 (1999).
27. Itsara, A. *et al. De novo* rates and selection of large copy number variation. *Genome Res.* **20**, 1469–1481 (2010).
28. Tian, Z. *et al.* Artificial selection for determinate growth habit in soybean. *Proc. Natl. Acad. Sci. USA* **107**, 8563–8568 (2010).
29. Zheng, P. *et al.* A phenylalanine in DGAT is a key determinant of oil content and composition in maize. *Nat. Genet.* **40**, 367–372 (2008).
30. Uauy, C., Distelfeld, A., Fahima, T., Blechl, A. & Dubcovsky, J. A NAC Gene regulating senescence improves grain protein, zinc, and iron content in wheat. *Science* **314**, 1298–1301 (2006).
31. Poland, J.A., Bradbury, P.J., Buckler, E.S. & Nelson, R.J. Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. *Proc. Natl. Acad. Sci. USA* **108**, 6893–6898 (2011).
32. Kump, K.L. *et al.* Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat. Genet.* **43**, 163–168 (2011).
33. Tester, M. & Langridge, P. Breeding technologies to increase crop production in a changing world. *Science* **327**, 818–822 (2010).
34. Moose, S.P. & Mumm, R.H. Molecular plant breeding as the foundation for 21st century crop improvement. *Plant Physiol.* **147**, 969–977 (2008).

ONLINE METHODS

Inbred line resequencing and variant identification. Total DNA from 278 maize inbred lines was extracted with the hexadecyltrimethylammonium bromide (CTAB) method for Illumina sequencing. Paired-end reads obtained from sequencing were mapped to the maize B73 genome with Burrows-Wheeler Aligner (BWA) software[35]. SAMtools[36] was used to convert mapping results to bam format, and duplicated reads were filtered with the help of the Picard package[36].

SNP detection was performed using the Genome Analysis Toolkit (GATK, version 1.0.4418)[37], as it supports multi-sample analysis. With the exception of the four inbred lines (8112, 5003, 478 and Zheng58) that were analyzed by deep resequencing whose SNPs were called separately as a group, our SNP pipeline aligned reads of at least 80 inbred lines all together with an overall coverage of greater than 200×, which significantly minimizes error in SNP calling. Realignment around indels was performed first to avoid alignment errors. Two steps of realignment were performed in GATK: the first step with the RealignerTargetCreator package identified regions in which realignment was needed, and the second step with IndelRealigner performed realignment within the regions found in the first step. After realignment, base-quality score recalibration was performed with two packages (CountCovariates and TableRecalibration). SNP calling was performed with UnifiedGenotyper, and mapping adjustment was then performed. The threshold of SNP calling was set to 20 for both base quality and mapping quality. As recommended by the GATK software, we set the confidence score of SNP calling to be more than 50, with the parameter -stand_call_conf set to 50. Four extra filtration steps were used for SNP calling in the four lines with deep-resequencing data (5003, 8112, 478 and Zheng58) with an average coverage of 27×. SNPs were discarded (i) if the mapping quality of 10% of the reads that covered a SNP were 0, (ii) if they had 10 bp around indels (excluded using GATK and a python script, makeIndelMask), (iii) if they occurred in a cluster (more than three SNPs within 10 bp) and (iv) if the coverage of the SNP locus was outside of 5–300×. All SNP annotation was performed according to the second version of the maize B73 genome.

SAMtools software was used for 1–10 nt indel detection with mapping quality set to ≥20. Only homozygous indels with more than five reads were recorded.

The method for the detection of CNVs was based on a described event-wise testing algorithm[38] with some adjustments. Read depth of every 100-bp window was computed by counting the start position of reads within this window. Considering the bias in read depth caused by GC content, we first adjusted the read depth of every window with the equation Adjusted_read Depth = readDepth × m / (m_{GC}), where Adjusted_readDepth is the adjusted read depth, readDepth is the read depth of the window, m is the median value of all windows of a chromosome and m_{GC} is the median read depth of all windows that have the same GC content as the adjusted window. After adjustment for GC content, we carried out CNV detection using the event-wise testing algorithm.

GWAS with the 278 inbred lines. We performed a GWAS for three traits: cob color, date to anthesis and silk color. All inbred lines were planted on 8 May 2011 in the experimental station of the China Agricultural University. Five randomly selected plants in the middle of the plots for each line were measured, and their mean value was used for the GWAS. The three traits were scored. Date to anthesis was measured as the days after planting to 50% anthesis and shed pollen at or near flowering time. Silk color was measured when the filaments were exposed from the ear by approximately 3 cm. After harvest, ears were dried naturally, and cob color was determined.

Because of the lack of HapMap reference, three available software programs (Beagle[39], Fastphase[40] and Npute[41]) were tested for imputation. To compare the accuracy of the three software programs, we randomly selected 1,000 SNPs and randomly missed 1 site to check whether the sequence from imputation was the same as that in the sequencing result. The same processes were carried out 1,000 times. The accuracy of Beagle, Fastphase and Npute were 95.2%, 92.1% and 93.5%, respectively, and Beagle used the least computation time. Therefore, we used Beagle to impute missing genotypes. Population structure was estimated with GCTA[42] tools. We used a compressed mixed linear model to perform the GWAS with GAPIT[43] software.

Validation of SNP calling through local *de novo* assembly. We chose 1,000 single-copy genes and conducted reference-guided local assembly using Schneeberger's pipeline with some modification[44]. Reads mapped to each gene were grouped together. There were three types of reads in each set: (i) paired-end reads with both reads mapped, (ii) paired-end reads with an unmapped read and (iii) single reads for which the pair mapped outside of gene regions. Each read set was assembled using CAP3 (ref. 45) (parameters of $-z = 2$, $-u = 2$, $-v = 2$, $-o = 20$, $-j = 35$, $-s = 251$ and $-h = 80$). Paired-end reads in each read set were used for scaffolding CAP3 contigs with SSAPCE[46] (parameters of $-t = 5$ and $-k = 2$). SSAPCE scaffolds were used for discovering variations. Genes with no more than five contigs were used to validate SNPs using BWA and SAMtools software. The consistency rate was calculated by comparing the SNPs called from short-read mapping with the local *de novo* assembly.

Simulation test on read alignment of Mo17 BAC sequences. All nine Mo17 BAC sequences were downloaded from NCBI. The BWA mapping tool generates mapping quality of 0 for repetitive sequences that are mapped to multiple sites. Because our pipeline used reads with mapping quality of greater than 20, we used only the non-repeat region (single-copy region, mapped uniquely in the genome) of the BAC sequences to generate 25× simulative reads by Mapping and Assembly with Qualities (MAQ). SNPs and indels were called by our pipeline using these reads, which were compared with the read obtained with the long BAC sequences in SAMtools.

Population genetics analysis and selective sweep scanning. Only SNPs with less than 50% missing were used for population analysis and selective sweep scanning. The neighbor-joining tree of the 126 US lines was constructed with PHYLIP[47] version 3.69. To avoid effects from population structure, we first compared each pair of the 278 inbred lines. Nucleotide diversity (π)[48], Tajima's D values[49] and F_{ST} values[50] were calculated in nonoverlapping windows of 200 SNPs using the libsequence C++ library[51] and in-house Perl scripts. Selective sweep signals were determined by calculating CLR as described[11]. A CLR test was calculated on each 50-kb window across each chromosome. Contiguous windows with 10% tails of CLR were merged. Genomic regions with the top 1% CLR values were considered to be targets of selection. Genes within selection targets were considered to be candidate selection genes.

Recombination map and mutation rates in pedigree inbred lines. We used a sliding-window method to construct the recombination map of the 478 and Zheng58 lines[52]. For 478, SNPs were filtered by two criteria: (i) the SNP site had to be sequenced no less than five times in all three inbred lines (5003, 8112 and 4780), and (ii) the SNP had to be polymorphic between the two parents (5003 and 8112). Sliding windows were used to calculate the SNP ratio between 8112 and 5003 along each chromosome, with a window size of 1,015 SNPs (~1 Mb of physical distance) and a step size of 105 SNPs. A breakpoint was defined when the SNP ratio switched from >1 to <1. For Zheng58, because there were only data from one parent, we used the SNP sites with sequencing depth of ≥5 in both 478 and Zheng58.

To investigate the mutation rate during breeding, we divided the genome into 5003 origin and 8112 origin. Windows with a parent SNP ratio of ≥90% that were 10 kb away from breakpoints were selected for further analysis. All SNPs in the resulting windows of 478 were compared again to its parents, 5003 and 8112. Windows in which more than 20% of SNPs were different from both parents were excluded. Regions covered by all three generations (5003, 478 and Zheng58, or 8112, 478 and Zheng58) were used for mutation rate calculations.

To identify high-confidence IBD regions for mutation rate calculation, windows with a parental SNP ratio of ≥90% that were 10 kb away from breakpoints were selected. Only regions shared by three generations (5003, 478 and Zheng58, or 8112, 478 and Zheng58) were selected for mutation rate calculation. Single-nucleotide mutation rate and the mutation rate of short indels of 1–10 bp were calculated as in *C. elegans*[25].

35. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
36. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

37. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
38. Yoon, S., Xuan, Z., Makarov, V., Ye, K. & Sebat, J. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* **19**, 1586–1592 (2009).
39. Browning, B.L. & Browning, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–223 (2009).
40. Stephens, M. & Scheet, P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.* **76**, 449–462 (2005).
41. Roberts, A. *et al.* Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. *Bioinformatics* **23**, i401–i407 (2007).
42. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
43. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).
44. Schneeberger, K. *et al.* Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. *Proc. Natl. Acad. Sci. USA* **108**, 10249–10254 (2011).
45. Huang, X. & Madan, A. CAP3: a DNA sequence assembly program. *Genome Res.* **9**, 868–877 (1999).
46. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
47. Felsenstein, J. PHYLIP: phylogeny inference package (version 3.2). *Cladistics* **5**, 164–166 (1989).
48. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
49. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
50. Hudson, R.R., Boos, D.D. & Kaplan, N.L. A statistical test for detecting geographic subdivision. *Mol. Biol. Evol.* **9**, 138–151 (1992).
51. Thornton, K. Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* **19**, 2325–2327 (2003).
52. Huang, X. *et al.* High-throughput genotyping by whole-genome resequencing. *Genome Res.* **19**, 1068–1076 (2009).

doi:10.1038/ng.2312