nature genetics

Genome-wide patterns of genetic variation among elite maize inbred lines

Jinsheng Lai[1,2,7], Ruiqiang Li[3,7], Xun Xu[3,7], Weiwei Jin[2,7], Mingliang Xu[2,7], Hainan Zhao[1,2], Zhongkai Xiang[1,2], Weibin Song[1,2], Kai Ying[4], Mei Zhang[1,2], Yinping Jiao[1,2], Peixiang Ni[3], Jianguo Zhang[3], Dong Li[3], Xiaosen Guo[3], Kaixiong Ye[3], Min Jian[3], Bo Wang[3], Huisong Zheng[3], Huiqing Liang[3], Xiuqing Zhang[3], Shoucai Wang[2], Shaojiang Chen[2], Jiansheng Li[2], Yan Fu[4], Nathan M Springer[5], Huanming Yang[3], Jian Wang[3], Jingrui Dai[2], Patrick S Schnable[4] & Jun Wang[3,6]

We have resequenced a group of six elite maize inbred lines, including the parents of the most productive commercial hybrid in China. This effort uncovered more than 1,000,000 SNPs, 30,000 indel polymorphisms and 101 low-sequence-diversity chromosomal intervals in the maize genome. We also identified several hundred complete genes that show presence/absence variation among these resequenced lines. We discuss the potential roles of complementation of presence/absence variations and other deleterious mutations in contributing to heterosis. High-density SNP and indel polymorphism markers reported here are expected to be a valuable resource for future genetic studies and the molecular breeding of this important crop.

The maize genome is large and complex[1–4]. Its genetic variation has been characterized by using molecular markers and by sequencing multiple alleles from selected loci[5,6]. With the advent of 'next generation' sequencing technology, it has become feasible to resequence entire large genomes[7,8] and thereby to carry out genome-wide surveys of genetic variation. The sequencing of the inbred B73 maize line[9] provides a reference genome that can be used to anchor resequencing data from other maize lines. Here, we analyze the whole-genome resequencing of six elite commercial maize inbred lines.

Inbred lines (Zheng58, 5003, 478, 178, Chang7-2 and Mo17) were selected on the basis of their agronomic importance and genetic relationships. Lines Zheng58, Chang7-2, 178 and Mo17 are all members of a popular heterotic group used in China (Mo17 is also a member of an important heterotic group used in the USA). Zheng58 and Chang7-2 are the parents of the commercial hybrid (ZD958) that is currently the most widely grown in China. Inbred line 178 is the female parent of another hybrid (ND108) that is also widely grown in China. Inbred line 478 is a parent and inbred line 5003 is a grandparent of Zheng58 (**Fig. 1**).

Resequencing yielded 1.26 billion 75-bp paired-end reads, which comprised 83.7 Gb of high-quality raw data. Sequence reads were aligned to the maize reference genome using SOAP software v2.18 (ref. 10). In total, we achieved an effective depth of ×32.4 coverage, with an average of ×5.4 for each inbred line (**Supplementary Table 1**).

We used unique reads in non-repeat regions to detect SNPs and indel polymorphisms (IDPs). SNPs were called with SOAPsnp[11] using a conservative quality filter pipeline (Online Methods). We identified 1,272,134 SNPs in non-repeat regions, with 468,966 in the 32,540 high-confidence maize genes (the 'filtered gene set') and 130,053 SNPs in coding regions. We also identified 30,178 indels ranging from 1 bp to 6 bp in length, of which 571 were in coding regions (**Supplementary Table 2**). Owing to the inherent relationships between some samples and to the characteristics of inbred lines, the

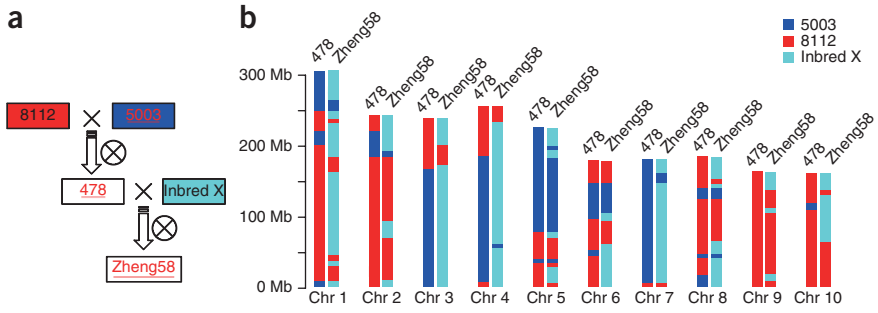

Figure 1 Genetic background of three sequenced inbred lines. (**a**) Pedigrees of three resequenced inbred lines. The female parent of each cross is listed first. The name of one parent of Zheng58 is not recorded and is termed 'Inbred X'. Resequenced inbred lines are underlined. (**b**) Reconstructed recombination events in inbred lines 478 and Zheng58 as they were derived from their parental lines.

[1]State Key Lab of Agrobiotechnology, China Agricultural University, Beijing, China. [2]National Maize Improvement Center, China Agricultural University, Beijing, China. [3]BGI-Shenzhen, Shenzhen, China. [4]Center for Plant Genomics, Iowa State University, Ames, Iowa, USA. [5]Department of Plant Biology, University of Minnesota, Saint Paul, Minnesota, USA. [6]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [7]These authors contributed equally to this work. Correspondence should be addressed to Jun Wang (wangj@genomics.cn), J. Lai (jlai@cau.edu.cn) or P.S.S. (schnable@iastate.edu).

Received 16 October 2009; accepted 13 September 2010; published online 24 October 2010; doi:10.1038/ng.684

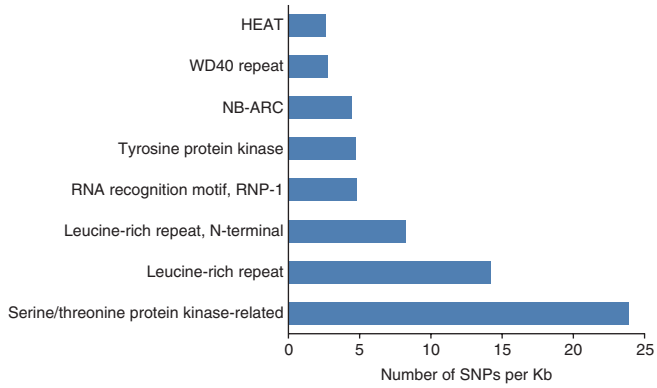

Figure 2 Annotation of large-effect SNPs. The numbers (shown by bar lengths) of large-effect SNPs for selected groups of genes are displayed. All the gene families shown were significantly abundant in large-effect SNPs (χ^2 test, $P < 0.01$).

overall genome diversity among these resequenced elite inbred lines (with a Watterson's θ of 0.0030; ref. 12) was lower than that reported for a more diverse population, which had a Watterson's θ of 0.0066 (ref. 13). Targeted sequencing of 92 randomly chosen PCR products validated 95% of the predicted SNPs. As three of the sequenced lines were closely related, cross-checking of SNPs identified from their chromosomal regions of common origin indicated that SNP accuracy was higher than 89%. This collection of SNPs and IDPs provides high-density marker coverage across the entire maize genome.

We also identified SNPs and IDPs that were predicted to have a potentially disabling effect on gene function: 1,478 SNPs were expected to induce premature stop codons, 97 were expected to alter initiation methionine residues, 828 were expected to disrupt splicing donor or acceptor sites, and 322 IDPs in coding regions were predicted to induce frame-shifts. In addition, 1,087 SNPs removed annotated stop codons, resulting in longer open reading frames. Of these large-effect SNPs, 101 were located in 46 genes that encoded disease-related proteins that contained leucine-rich repeat regions (**Fig. 2**), consistent with findings in *Arabidopsis*[14] and other plants[15,16]. As the identification of large-effect SNPs depends on the annotation of gene models, the exact number of such SNPs will probably be modified when the genome annotation is updated.

By projecting the SNPs onto all of the gene models (including 1 kb upstream of transcriptional start sites) in the filtered gene set, we identified 393 genes that contained no SNPs among the resequenced lines. To identify chromosomal regions that had low sequence diversity, we calculated the number of segregating nucleotide sites per site (K)[12] within 1-Mb sliding windows across the genome. A histogram of these data revealed a bimodal distribution with modal values of ~0.0020 and ~0.0068 SNPs per site (**Supplementary Fig. 1**). The distribution of the number of SNPs for each window fit well with a normal mixture model that has two components that define two classes of windows (**Supplementary Fig. 1**). We calculated the probability that any particular window belonged in a specific class (Online Methods) and identified 101 genomic intervals scattered throughout the genome that had reduced numbers of SNPs (**Supplementary Table 3**). These blocks had an average length of 2.4 Mb, with the longest (on chromosome 5) being ~13 Mb (**Fig. 3**, **Supplementary Fig. 2** and **Supplementary Table 3**). In these regions of low sequence diversity, we identified a number of genes that were known to have been under selection during maize improvement, including *bt2* and *su1* (ref. 17), 20% (6 out of 30) of a set of candidate selected genes[18], and ~43% (170 out of 393) of the zero-sequence-diversity genes identified here. Owing to the small sampling size and the relatedness of these resequenced inbred lines, the origin of these low diversity regions cannot be determined. They could have arisen through identity by descent (IBD) or through other types of demographic event. Many of these low-sequence-diversity regions overlapped with the low-recombination regions identified in the more diverse population[13], particularly for those in the pericentromeric regions.

Presence/absence variations (PAVs) have been described in maize genes[19]. However, these PAV genes were subsequently shown to be chimeric gene fragments carried by *Helitron* transposons[20,21]. A more recent study using comparative genome hybridization (CGH)[22] showed that hundreds of intact, single-copy, expressed genes that were present in the B73 genome were absent from the Mo17 genome.

To investigate the prevalence of PAVs, we mapped Mo17 resequencing reads to the B73 reference genome and identified 104 regions in the B73 genome in which at least 80% of a 5-kb or longer genomic region and at least 90% of its annotated transcriptional region (from the transcriptional start to stop) did not have corresponding Mo17 reads. These 104 regions were regarded as putative missing genes in the Mo17 inbred line. To rule out the possibility that some of these putative PAVs (pPAVs) reflected incomplete sampling, we compared our pPAVs to the existing CGH data[22]. High concurrence between the CGH data and the resequencing results indicated that most of the pPAVs reflected true differences in gene content between the B73 and Mo17 genomes.

We used the same criteria to identify PAVs in the other inbred lines, and found 296 high-confidence genes in B73 that were missing from at least one of the six inbred lines. Because the filtered gene set used was highly enriched for expressed genes and depleted for gene fragments and transposons, most of these PAVs represented intact, expressed genes. As expected from their pedigrees, three lines (Zheng58, 478 and 5003) from the same heterotic group had many deleted genes in common. **Figure 4** shows the number of PAV genes in each of the four heterotic groups (**Supplementary Table 4** gives a complete list of PAVs identified). Whereas most of the PAV events seemed to involve only a single gene, some included deletions of two to four adjacent genes. One large deletion on chromosome 6 of the Mo17 genome, which spans ~2 Mb with 24 genes annotated on the B73 reference genome, has at least 18 genes deleted according to our conservative criteria. It is possible that the entire region may have been deleted in Mo17, consistent with the results of CGH experiments[22].

To identify genes that were absent from B73 but present in the other six inbred lines, we assembled all non-mapping reads combined from all six lines using SOAPdenovo[23] and obtained contigs with a total

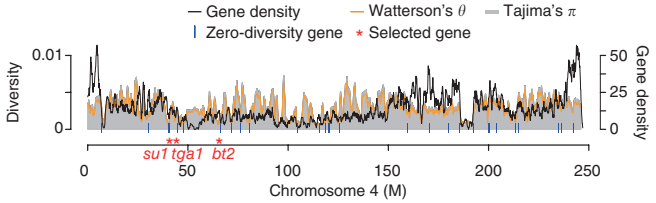

Figure 3 Genome-wide distribution of sequence diversity level, gene density, zero-diversity genes and selected genes on chromosome 4. Other chromosomes are shown in **Supplementary Figure 2**. Sequence diversity level (Watterson's θ per site[12], yellow line; Tajima's π per site[28], gray shading) and gene density (black line) are plotted using 1-Mb sliding windows. Regions of low genetic diversity are highlighted in green. Zero-diversity genes are shown by vertical blue bars. Red asterisks mark the positions of three genes known to show evidence of selection.

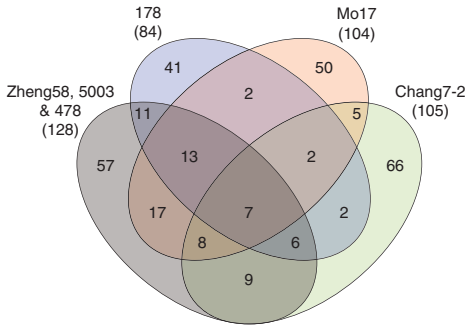

Figure 4 Numbers of PAVs relative to the B73 reference genome. Inbred lines Zheng58, 5003 and 478 were pooled for this analysis because they are members of the same heterotic group. The remaining three inbred lines were from three other heterotic groups. The numbers of PAVs in each of the four heterotic groups sampled in this study are shown in parentheses.

length of 5.4 Mb of low-copy sequences. Annotation of these contigs resulted in 570 putative absent genes (**Supplementary Table 5**) with an average length of 527 bp (considering only coding regions). Because these genes showed the same overall depth of sequencing coverage as other annotated genes (**Supplementary Table 5**), it seems unlikely that these reads were due to contaminant DNA. A Blast search against plant gene databases showed that 292 (55%) of these genes showed homology with plant proteins (**Supplementary Table 5**). Nearly half of these genes (267 out of 570; 47%) could be functionally classified by InterPro[24]. Seven were members of a leucine-rich repeat family and two belonged to the NB-ARC (nucleotide-binding adaptor shared by R proteins) family, indicating that they might be involved in strain-specific disease resistance (**Supplementary Table 5**).

Because some of these newly identified genes might exist in the B73 genome, but were not included in the current B73 assembly[12], we resequenced the B73 inbred line using Solexa 75-bp pair-end reads. This resequencing showed with high confidence that at least 157 of these 572 putative genic contigs were missing from the B73 genome because they had no corresponding reads in the 20× whole-genome resequencing dataset (**Supplementary Table 5**). About 300 of these putative genic contigs had B73 resequencing reads that covered more than 90% of the assembled contigs (**Supplementary Table 5**), suggesting that these putative genes were present in B73 but were missed in the current genome release[12].

PAVs and deleterious mutations, such as the large-effect SNP and IDPs described here, offer an opportunity to test hypotheses of heterosis (also known as hybrid vigor), a phenomenon that has been found in many species but has not been mechanistically explained[19,25]. The dominance hypothesis proposes that heterosis is the result of complementation of slightly deleterious recessive alleles and that the fixation of these alleles in inbred lines results in inbreeding depression[19,25,26]. It is reasonable to assume that some of the identified PAV genes are functional because most of them are expressed. Inbred lines that have large differences in gene content could therefore complement one another, contributing to heterosis. We assessed the differences in the gene sets of different heterotic groups from our resequenced lines. Consistent with this hypothesis, **Figure 4** shows that inbred lines representing different heterotic groups primarily contain different sets of deleted genes. For example, few deleted genes are shared between the Zheng58/478/5003 group and their heterotic partner, Chang7-2. In addition, only 6% of genes that are putatively absent from B73 have paralogs (>80% nucleotide identity and >80% length) in the B73 reference genome, suggesting that most of these missing genes are not redundant in function. In addition to PAVs, other deleterious mutations, including SNPs and IDPs that have a large and potentially disabling impact, could also potentially contribute to heterosis. Only 33% of the 2,033 B73 genes that contained large-effect SNPs had paralogs in the B73 reference genome, consistent with the complementation hypothesis. However, it is unlikely that heterosis is the result of any single mechanism[25]. Observed levels of residual heterozygosity around pericentromeric regions in recombinant inbred lines of the Nested Association Mapping (NAM) population were recently used to support of the idea that pseudo-overdominance contributes to heterosis[13,27]. Similarly, repulsion-phase linkage of PAVs and deleterious mutations could also contribute to pseudo-overdominance of heterosis.

This resequencing project was designed to allow us to evaluate the genomic changes that occurred during pedigree breeding (**Supplementary Fig. 3**). The high-density SNP polymorphisms discovered and the known pedigree relationships allowed us to reconstruct, at high resolution, the recombination events that gave rise to specific inbred lines. We tracked chromosomal segments through two cycles of pedigree breeding involving inbred lines 5003, 478 and Zheng58. Following the pedigree in **Figure 1a**, our analysis showed that there were 27 recombination breakpoints from inbred line 5003 to inbred line 478 and 46 breakpoints from inbred line 478 to inbred line Zheng58. Inbred line 478 inherited 43% of its genome from one parent (5003) and 57% from its other parent (8112). By contrast, Zheng58 inherited 43% of its genomic content from inbred line 478, but the contributions from its grandparents, inbred lines 5003 and 8112, were unequal (12% from 5003 and 31% from 8112; **Fig. 1b**).

We report here a whole genome map of SNPs, IDPs and gene content variation among elite maize lines. Our results suggest that gene content complementation might be an important factor in heterosis in maize. However, additional evidence obtained through analysis of more inbred lines will be required to establish the strength of this hypothesis and the potential contribution of this mechanism. Genome-wide association studies with yield data of hybrids generated from various inbred combinations using the SNPs discovered here will provide additional clues to the molecular basis of heterosis and will help researchers to identify quantitative trait loci that are important for crop improvement.

URLs. Detailed information on novel sequences, novel gene annotation and IDPs is available at ftp://rice:ricedownload@public.genomics.org.cn/BGI/maize. SOAP and SOAPsnp are available at http://soap.genomics.org.cn/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. The sequence data have been deposited in NCBI Short Read Archive with accession number SRA010130. The whole genome SNP dataset has been deposited in NCBI dbSNP with accession number records ss181800510–ss184955572.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

Supported by the 973 program (2009CB118400; 2007CB815703; 2007CB815705; 2007CB109000), the 863 project (2010AA10A106), the National Natural Science Foundation of China (30725008), the Shenzhen Bureau of Science Technology & Information, China (ZYC200903240077A; CXB200903110066A), the Chinese Academy of Science (GJHZ0701-6), the Ole Rømer grant from the Danish

Natural Science Research Council and the US National Science Foundation (DBI-0527192). We thank L. Goodman for editing the manuscript.

AUTHOR CONTRIBUTIONS

J. Lai, Jun Wang, R.L., J.D. and P.S.S. managed the project. X.X., H. Zhao, Z.X., W.S., M.Z., Y.J., P.N., M.J., B.W., H. Zheng, H.L. and X.Z. performed experiments and sequencing. J. Lai, Jun Wang, R.L., X.X., Jian Wang and H.Y. designed the analyses. X.X., R.L., W.J., M.X., K. Ying, J.Z., D.L., X.G., K. Ye, S.W., S.C., J. Li and Y.F. performed data analyses. J. Lai, P.S.S., N.M.S., Jun Wang, K. Ying and X.X. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. SanMiguel, P. *et al.* Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274**, 765–768 (1996).
2. Messing, J. *et al.* Sequence composition and genome organization of maize. *Proc. Natl. Acad. Sci. USA* **101**, 14349–14354 (2004).
3. Whitelaw, C.A. *et al.* Enrichment of gene-coding sequences in maize by genome filtration. *Science* **302**, 2118–2120 (2003).
4. Palmer, L.E. *et al.* Maize genome sequencing by methylation filtration. *Science* **302**, 2115–2117 (2003).
5. Tenaillon, M.I. *et al.* Patterns of diversity and recombination along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Genetics* **162**, 1401–1413 (2002).
6. Tenaillon, M.I. *et al.* Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proc. Natl. Acad. Sci. USA* **98**, 9161–9166 (2001).
7. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
8. Wheeler, D.A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
9. Schnable, P.S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
10. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713–714 (2008).
11. Li, R. *et al.* SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19**, 1124–1132 (2009).
12. Watterson, G.A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
13. Gore, M.A. *et al.* A first-generation haplotype map of maize. *Science* **326**, 1115–1117 (2009).
14. Clark, R.M. *et al.* Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* **317**, 338–342 (2007).
15. Bakker, E.G., Toomajian, C., Kreitman, M. & Bergelson, J. A genome-wide survey of R gene polymorphisms in *Arabidopsis*. *Plant Cell* **18**, 1803–1818 (2006).
16. Grant, M.R. *et al.* Independent deletions of a pathogen-resistance gene in *Brassica* and *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* **95**, 15843–15848 (1998).
17. Whitt, S.R., Wilson, L.M., Tenaillon, M.I., Gaut, B.S. & Buckler, E.S. IV. Genetic diversity and selection in the maize starch pathway. *Proc. Natl. Acad. Sci. USA* **99**, 12959–12962 (2002).
18. Wright, S.I. *et al.* The effects of artificial selection on the maize genome. *Science* **308**, 1310–1314 (2005).
19. Fu, H. & Dooner, H.K. Intraspecific violation of genetic colinearity and its implications in maize. *Proc. Natl. Acad. Sci. USA* **99**, 9573–9578 (2002).
20. Lai, J., Li, Y., Messing, J. & Dooner, H.K. Gene movement by Helitron transposons contributes to the haplotype variability of maize. *Proc. Natl. Acad. Sci. USA* **102**, 9068–9073 (2005).
21. Morgante, M. *et al.* Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat. Genet.* **37**, 997–1002 (2005).
22. Springer, N.M. *et al.* Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet.* **5**, e1000734 (2009).
23. Li, R. *et al. De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2009).
24. Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116–120 (2005).
25. Springer, N.M. & Stupar, R.M. Allelic variation and heterosis in maize: how do two halves make more than a whole? *Genome Res.* **17**, 264–275 (2007).
26. Charlesworth, D. & Willis, J.H. The genetics of inbreeding depression. *Nat. Rev. Genet.* **10**, 783–796 (2009).
27. McMullen, M.D. *et al.* Genetic properties of the maize nested association mapping population. *Science* **325**, 737–740 (2009).
28. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).

ONLINE METHODS

SNP detection. We used a four-step procedure to detect high-quality SNPs[29]. (i) We calculated the likelihood of each accession's genotype using SOAPsnp[11]. (ii) On the basis of the resequencing data of six accessions, sites with sufficient quality, called effective sites, were used for SNP determination. Sufficient quality was based on the following criteria: 10 ≤ total depth ≤ 100, total depth calculated by combining data from all six individuals, average mappable sites < 1.5. All individual likelihood files were integrated to produce a pseudo-genome for each site of the total sample using maximum likelihood estimation (MLE). Sites that passed the criteria according to read mapping uniqueness (average mappable sites < 1.5, which means most reads that cross a site can find unique hits in the reference genome), sequencing depth (10~100 for the total 6 inbred lines), quality score (average quality for the novel allele >20) and minor allele count (supported by ≥5 reads) were kept as candidate SNPs. To exclude SNP calling errors caused by incorrect mapping or indels, we did not call two adjacent SNPs that were separated by <5 bp. The remaining SNPs were defined as high-quality (HQ) SNPs. (iii) We performed SNP calling for the six inbred lines together. SNPs for each individual inbred line comprised a subset of the total HQ SNP set. (iv) Base types were allocated back to each individual on the basis of the genotypes of the HQ SNPs and each individual likelihood file. The genotype with the largest likelihood was chosen as the consensus genotype for each individual. These SNPs were used to calculate the whole genome SNP number and the diversity pattern of the whole genome.

SNP annotation. The localization of SNPs in coding regions, non-coding regions, start codons, stop codons and splice sites was based on annotation of gene models as provided by the Maize Genome Sequencing Project website (see URLs). The identification of synonymous and non-synonymous status for SNPs within the coding sequence (CDS) was conducted using Genewise software[30]. To functionally annotate each gene, we first selected the longest mRNA for each gene, and then aligned the corresponding protein sequence translated from its CDS to Pfam using InterProScan[24]. Of the 32,540 non-redundant proteins, 11,150 had at least one best hit and the functional annotation for each gene was obtained from its corresponding InterPro entry, when applicable.

Identification of low-polymorphic chromosomal regions and zero-diversity genes. Sequence diversity was estimated using a commonly used measure of DNA polymorphism, the number of segregating nucleotide sites (K) per site[12] in a sample of sequences, where a segregating site is a site that shows variation among the sequences in the sample. To calculate sequence diversity across the genome, a sliding window method was used to analyze each chromosome separately with a window size of 1 Mb and a sliding step of 100 kb. Windows with insufficient effective length (total number of effective sites less than 10% of the window length) were not used for sequence diversity calculations.

To better characterize the bimodal distribution, a two-component normal mixture model[31] was used to simulate the distribution of the sequence densities. The expectation-maximization (EM) algorithm[32] was used to estimate the mixing proportion, the mean and the variance of each component. On the basis of this simulation one component had a mean near 0.0021 (class I, low diversity region) and the other had a mean near 0.0068 (class II, normal diversity region). The simulation also provides the probability with which each window belongs to class I (component I > 0.75), class II (component II > 0.75) or is uncertain.

Reconstructing the pedigree breeding history. We scanned the diversity-decreased level $((D_{arv} - D_{pair})/D_{arv}$, where D_{arv} is an estimate of the average diversity level for all six samples in one region in the reference and D_{pair} is an estimate of the diversity level of two samples) for each parent and descendant pairs with sliding windows. Bimodal distributions, which represent the haplotype blocks that are descended from different parents, were identified, and were assessed as descending from a particular parental strain. Uncertain regions (on the boundary between two types) were equally apportioned into the two adjacent blocks. We then set different colors for genomic regions according to the type of haplotype block (**Fig. 1b**).

29. Xia, Q. *et al.* Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (Bombyx). *Science* **326**, 433–436 (2009).
30. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
31. Everitt, B.S. An introduction to finite mixture distributions. *Stat. Methods Med. Res.* **5**, 107–127 (1996).
32. Dempster, A.P., Laird, N.M. & Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. (Ser. A)* **39**, 1–38 (1977).