

Maize HapMap2 identifies extant variation from a genome in flux

Jer-Ming Chia^{1,24}, Chi Song^{2,24}, Peter J Bradbury^{3,4}, Denise Costich^{3,4}, Natalia de Leon^{5,6}, John Doebley⁷, Robert J Elshire³, Brandon Gaut⁸, Laura Geller¹, Jeffrey C Glaubitz³, Michael Gore^{4,9}, Kate E Guill^{4,10}, Jim Holland^{4,11}, Matthew B Hufford¹², Jinsheng Lai^{13–15}, Meng Li³, Xin Liu², Yanli Lu¹⁶, Richard McCombie¹, Rebecca Nelson¹⁷, Jesse Poland^{4,18}, Boddupalli M Prasanna¹⁹, Tanja Pyhäjärvi¹², Tingzhao Rong¹⁶, Rajandeep S Sekhon^{5,6}, Qi Sun³, Maud I Tenaillon^{8,20}, Feng Tian³, Jun Wang², Xun Xu², Zhiwu Zhang³, Shawn M Kaeppeler^{5,6}, Jeffrey Ross-Ibarra^{12,21}, Michael D McMullen^{4,10}, Edward S Buckler^{3,4}, Gengyun Zhang², Yunbi Xu^{22,23} & Doreen Ware^{1,4}

Whereas breeders have exploited diversity in maize for yield improvements, there has been limited progress in using beneficial alleles in undomesticated varieties. Characterizing standing variation in this complex genome has been challenging, with only a small fraction of it described to date. Using a population genetics scoring model, we identified 55 million SNPs in 103 lines across pre-domestication and domesticated *Zea mays* varieties, including a representative from the sister genus *Tripsacum*. We find that structural variations are pervasive in the *Z. mays* genome and are enriched at loci associated with important traits. By investigating the drivers of genome size variation, we find that the larger *Tripsacum* genome can be explained by transposable element abundance rather than an allopolyploid origin. In contrast, intraspecific genome size variation seems to be controlled by chromosomal knob content. There is tremendous overlap in key gene content in maize and *Tripsacum*, suggesting that adaptations from *Tripsacum* (for example, perennialism and frost and drought tolerance) can likely be integrated into maize.

Maize nucleotide diversity is near the upper limit of that estimated for crops and is an order of magnitude higher than in humans^{1,2}. However, this genetic diversity not only arises from SNPs and

small insertions-deletions (indels) but also from larger structural variations^{3–5}. Additionally, homeologous segments in this paleopolyploid contribute to genome complexity. Nearly 85% of the maize reference genome sequence is annotated as transposable elements^{6,7}, and cycles of transposable element invasion, activity and loss⁸, combined with the ability of these elements to shuffle gene fragments, have undoubtedly left a profound impact on the genome. Recent estimates suggest that the reference B73 sequence may capture only ~70% of the low-copy gene fraction of maize inbred lines⁹, with both genes and transposable elements occupying the unshared sequence space¹⁰. Consequently, in addition to SNPs and small indels, structural variations in the form of copy-number variations, presence/absence variations and movement of transposable elements^{3–5} contribute substantially to the genetic diversity in maize. Characterizing this diversity—in particular, at a sufficient density to drive genome-wide association studies (GWAS) or genomic selection—has been technically challenging.

Here, through whole-genome surveys using sequencing-by-synthesis technology, we have conducted a comprehensive characterization of genetic variation across 103 inbred lines representing a wide breadth of the *Z. mays* lineage, comprising 60 improved maize lines, including the parents of the maize nested association mapping (NAM) population¹¹, 23 maize landraces and 19 wild relatives (17 *Z. mays* ssp. *parviglumis* and 2 *Z. mays* ssp. *mexicana*)

¹Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. ²BGI-Shenzhen, Shenzhen, China. ³Institute for Genomic Diversity, Cornell University, Ithaca, New York, USA. ⁴US Department of Agriculture–Agriculture Research Service (USDA-ARS). ⁵US Department of Energy (DOE) Great Lakes Bioenergy Research Center, University of Wisconsin, Madison, Wisconsin, USA. ⁶Department of Agronomy, University of Wisconsin, Madison, Wisconsin, USA. ⁷Department of Genetics, University of Wisconsin, Madison, Wisconsin, USA. ⁸Department of Ecology and Evolutionary Biology, University of California, Irvine, California, USA. ⁹US Arid-Land Agricultural Research Center, Maricopa, Arizona, USA. ¹⁰Division of Plant Sciences, University of Missouri, Columbia, Missouri, USA. ¹¹Department of Crop Science, North Carolina State University, Raleigh, North Carolina, USA. ¹²Department of Plant Sciences, University of California, Davis, California, USA. ¹³State Key Laboratory of Agrobiotechnology, China Agricultural University, Beijing, China. ¹⁴National Maize Improvement Center, China Agricultural University, Beijing, China. ¹⁵Department of Plant Genetics and Breeding, China Agricultural University, Beijing, China. ¹⁶Maize Research Institute, Sichuan Agricultural University, Chengdu, China. ¹⁷Department of Plant Pathology and Plant Microbe Biology, Cornell University, Ithaca, New York, USA. ¹⁸Kansas State University, Manhattan, Kansas, USA. ¹⁹International Maize and Wheat Improvement Center (CIMMYT), Nairobi, Kenya. ²⁰Unité Mixte de Recherche (UMR) de Génétique Végétale, Institut National de la Recherche Agronomique (INRA)–Centre National de la Recherche Scientifique (CNRS)–Université Paris–Sud–AgroParisTech, Gif-sur-Yvette, France. ²¹The Genome Center, University of California, Davis, California, USA. ²²Institute of Crop Science, National Key Facilities for Crop Genetic Resources and Improvement, Chinese Academy of Agricultural Sciences, Beijing, China. ²³CIMMYT, Texcoco, Mexico. ²⁴These authors contributed equally to this work. Correspondence should be addressed to E.S.B. (esb33@cornell.edu), G.Z. (zhanggengyun@genomics.org.cn), M.D.M. (mcmullenm@missouri.edu), Y.X. (y.xu@cgiar.org) or D.W. (ware@cshl.edu).

Received 29 November 2011; accepted 7 May 2012; published online 3 June 2012; doi:10.1038/ng.2313

(Supplementary Table 1). We also generated sequence for a representative of the sister genus *Tripsacum*, *T. dactyloides* (Eastern gamagrass), as an outgroup for studying the evolutionary history of maize.

The nearly 1 trillion basepairs of DNA sequence generated from 13 billion reads were aligned against the B73 reference genome (RefGenV1)⁷, representing ~4.2× coverage of each of the maize and teosinte genomes and ~8× coverage of the *Tripsacum* genome. Two complementary algorithms were used for initial variant discovery. The first implements a Needleman-Wunsch–based alignment method that provides more sensitivity in identifying indels⁹, whereas the second better models sequencing errors¹². These algorithms, which separately perform well in human genetics¹³, failed to produce consistent results in maize—only 33% of the 93 million total segregating sites identified were shared between the two approaches (Supplementary Table 2). Taking advantage of the relative strengths of each approach, we applied a novel population genetics–based quality control pipeline to the combined set of variants. Briefly, we developed a logistic regression model based on patterns of linkage disequilibrium (LD) and both allele and genotype segregation. The model was trained on large regions of identity by descent (IBD) found using an independent SNP assay (Supplementary Table 3). Analysis of the final SNP set confirmed that the majority of the initial calls were likely due to paralogy (Supplementary Table 2). Overall, the final HapMap2 data set consisted of 55 million SNPs, with a 1% total error rate and a 5.6% minor allele error rate (Fig. 1, Supplementary Figs. 1 and 2 and Supplementary Table 2). The allele distribution of the calls across each line is listed (Supplementary Table 4). It is important to note, however, that many of the rejected variants are real, encoded from homeologous regions fostered by transposable element invasion and loss, a fact that underscores the limitations of relying on a single reference genome.

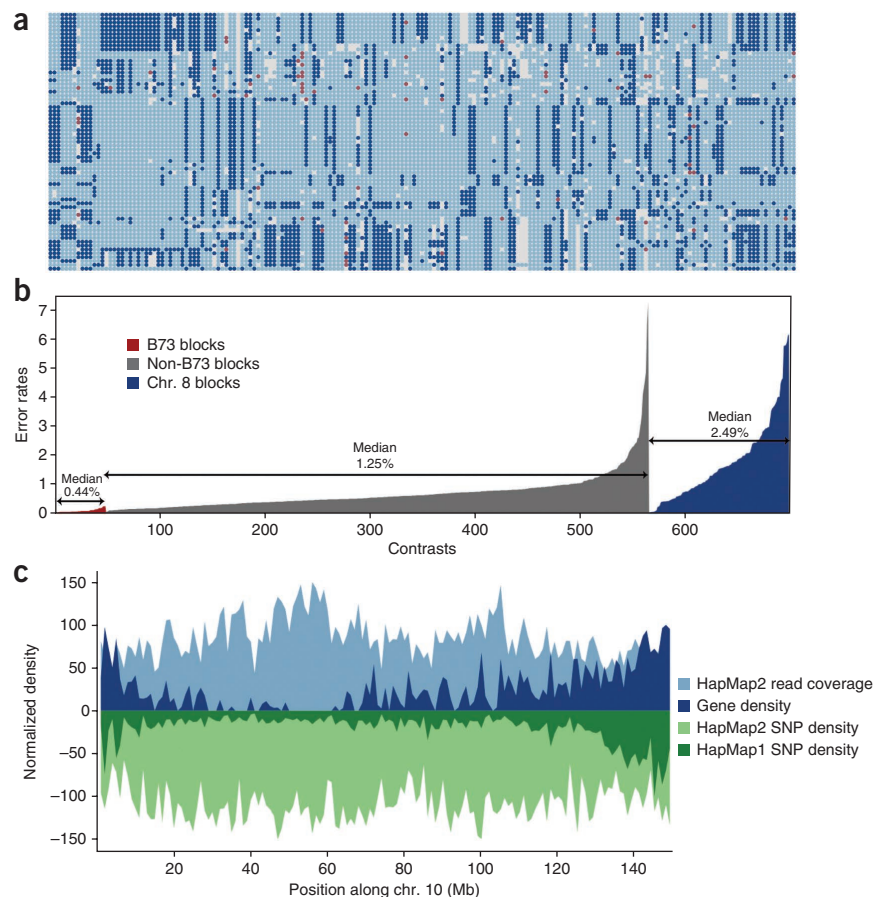
We examined the 55 million SNPs for their potential effects on protein-coding sequences. We found that 21% were associated with a genic region, including 825,000 synonymous, 571,000 nonsynonymous and ~10,000 nonsense mutations (Supplementary Fig. 3). More than 1,500 (7.5%) of the 20,380 high-confidence genes (those with full-length cDNA support) carried a premature stop codon variation. The allele frequency of these nonsense SNPs differed between the

improved maize lines and teosinte (8.0% versus 9.6%; Supplementary Table 5), and teosinte genomes carried more nonsense mutations on average (Supplementary Table 6). This lower genetic load in maize inbred lines relative to their outcrossing ancestor teosinte is consistent with the hypothesis that homozygosity purges recessive deleterious mutations, as seen in selfing taxa, such as *Arabidopsis*¹⁴.

We characterized structural variation in the maize genome through a global analysis of read-depth variants (RDVs) in both 10-kb windows and individual genes. Our data suggest that the entire maize genome is in flux: more than 90% of the 10-kb windows showed greater than twofold variation in read depth at a false discovery rate of 1%, and more than 70% of windows had such RDVs in ≥10 lines. By comparing the RDVs to nearby SNPs using an LD test, 80% of the tested RDV intervals could be anchored locally (Supplementary Note). The majority (70%) of genes had an RDV in at least one line, and nearly a third (32%) had RDVs in ≥10 lines (Supplementary Table 7). Notably, as with tandemly arrayed genes in rice and *Arabidopsis*¹⁵, genes with high levels of RDVs were found more often in gene ontology (GO) categories of stress and stimulus responses, whereas structurally invariant genes more often encoded constituent biological processes (Supplementary Table 8).

Despite the tremendous amount of historic recombination that has occurred in *Z. mays*^{1,2}, we found that large haplotype blocks were nonetheless evident throughout the genome. In the maize lines studied here, LD was generally low and decayed to an average $r^2 = 0.2$ in 5,500 bp (Fig. 1), but there were still extensive haplotypes shared among improved lines. Across all of the maize lines we analyzed, we found 80 blocks of IBD larger than 10 Mb in size (Supplementary Table 3), which is consistent with the results

Figure 1 Deriving a high-quality variation map from a fluid genome. (a) IBD blocks in a 9-Mb pericentromeric region of chromosome 10 were used to develop a model for scoring variation. Each column represents a SNP, and each row is a haplotype from an inbred line. Dark and light blue circles represent the major and minor alleles, respectively, at each site. Gray spaces represent missing genotypes, and red circles represent heterozygous cells. (b) Error rates based on IBD estimates broken down by individual contrasts between IBD pairs. Contrasts involving B73 had almost threefold lower error rates than the median. IBD blocks on chromosome 8 had higher rates, likely due to errors in the reference sequence or assignment of IBD. (c) Unlike in the previous iteration of maize HapMap⁹, read coverage and variation density were not restricted to genic regions in the present analysis. The plot shows the distribution of SNP and gene density in HapMap1 and HapMap2.



from haplotype analysis of a smaller sample of Chinese inbred maize lines¹⁶. This lack of recent recombination in some regions of the genome is likely central to the pseudo-overdominance model of heterosis^{3,5,9,11}. Considering teosinte and maize lines together reduces the decay of LD to only 1 kb (Supplementary Fig. 4), which is a result that helps justify the development of high-resolution association mapping populations that include teosinte. Notably, in contrast to HapMap1, in which the SNPs were in high LD in only 34% of pairwise comparisons⁹, the variants described here were closer to saturating the genome with polymorphic markers in tight LD (80% of pairwise comparisons; $r^2 > 0.8$), finally allowing genome-wide association studies to be performed in maize.

We evaluated the usefulness of our data for GWAS by combining the SNPs and RDVs identified here with the 1.6 million SNPs from HapMap1 (ref. 9) in an association analysis of 5 key traits involved in leaf development and disease resistance^{17–19}. Overall, we found better agreement of the complete marker set with linkage mapping peaks than with HapMap1 SNPs alone (Supplementary Fig. 5); in many cases, associations were much stronger with the complete set (Supplementary Table 9). HapMap2 SNPs contributed most to significantly associated loci (66%), and, in terms of marker types, genic SNPs (from both HapMap1 and 2), comprised over 60% of significant markers, with RDVs comprising 7% (Fig. 2). Unexpectedly, RDVs were overrepresented in the GWAS results, even after taking into account their abundance in the genome, with genic and 10-kb RDVs enriched up to 11- or 18-fold, respectively (Fig. 2). Given that LD in maize decayed to $r^2 = 0.2$ in 5.5 kb, we evaluated whether enrichment of RDVs at associated loci was present if smaller window sizes were

used. Indeed, we still observe enrichment of RDVs in loci associated with traits when using smaller window sizes, and, notably, although 2-kb RDVs made up only 3.5% of markers used in these tests, they contributed to 15–27% of associated loci (Supplementary Fig. 6 and Supplementary Table 10). This suggests that structural variation, captured here by RDVs, may have an important role in phenotypic variation. Furthermore, in species where complete reference assemblies are unavailable, RDVs of *de novo* contigs could be economical proxies for capturing structural variation, complementing SNP information derived from the same primary data.

Structural variation due to transposable element expansion²⁰ and variation in repeat arrays^{21–23} has previously been suggested to underlie genome size variation among maize accessions. We compared the abundance of knob repeats and more than 1,300 transposable element families to flow cytometry estimates of genome size for 38 lines (27 maize and 11 teosinte; Supplementary Fig. 7). We found that larger genomes were not associated with increased transposable element abundance, but genome size was positively correlated with the abundance of total knob repeats (Fig. 3). Except for the relative counts of transposable elements in the RLX_osed, RLX_sela and RLX_sari families ($r = 0.77, 0.78$ and 0.62 , respectively), the majority of the most abundant families were negatively correlated with genome size (Supplementary Tables 11 and 12). Previous work has indicated that transposable elements in the RLX_osed, RLX_sela and RLX_sari families are likely satellite repeats²⁴, and these elements correlated nearly perfectly with the abundance of knob repeats ($r = 0.98, 0.97$ and 1.00 , respectively; Supplementary Tables 13 and 14). Hence, whereas transposable elements are well

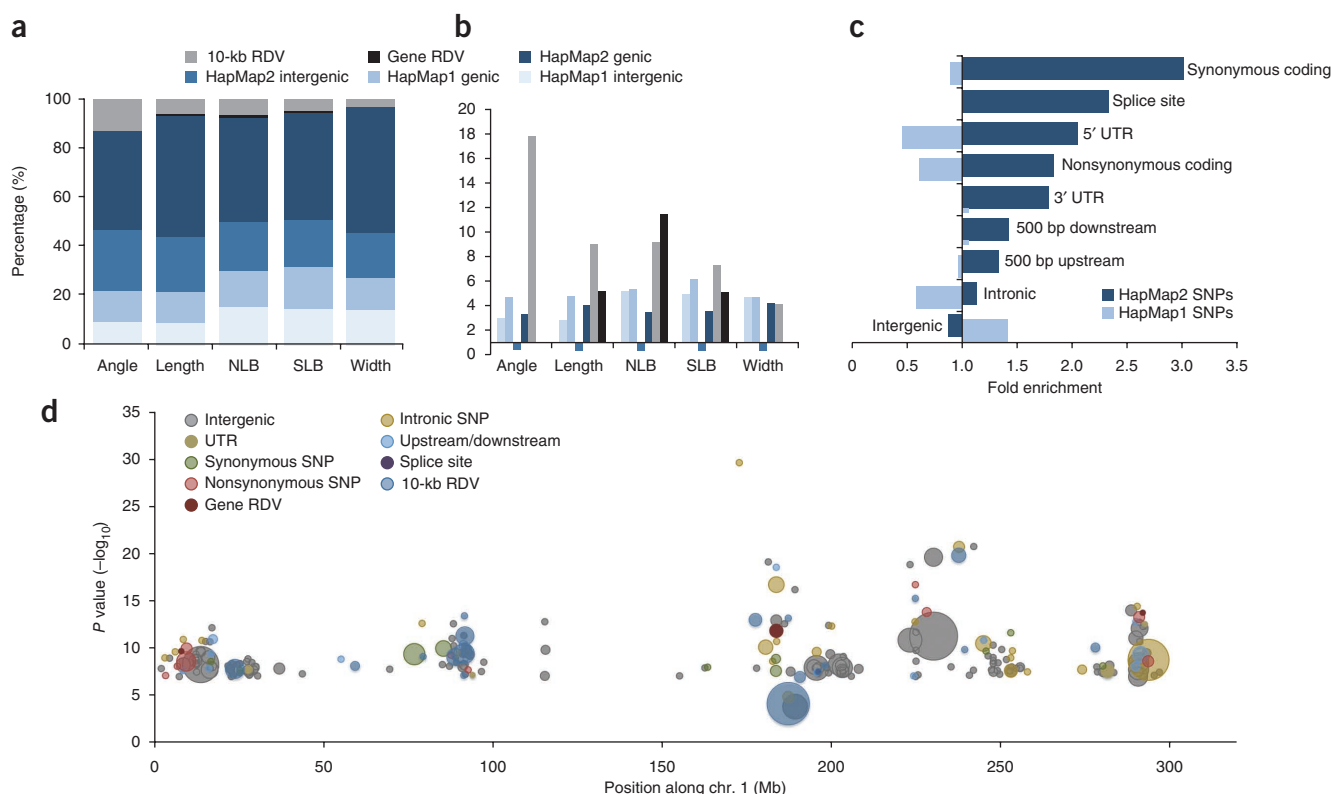


Figure 2 The impact of SNPs and RDVs on phenotype. GWAS on five traits: leaf angle, length and width and resistance to northern leaf blight (NLB) and southern leaf blight (SLB)^{17–19}. (a) The contribution of each class of markers to terms with bootstrap posterior probability (BPP) of ≥ 0.05 , showing that SNPs in HapMap2 contributed up to 66% of significantly associated markers. (b) Enrichment of each class of marker to terms with BPP of ≥ 0.05 . RDVs contributed disproportionately more to significant terms. (c) Enrichment of genic SNPs in terms with BPP of ≥ 0.05 . (d) Significant markers for NLB on chromosome 1. The vertical axis represents the log-scaled P value for each marker, and the size of each circle represents the BPP of the marker.

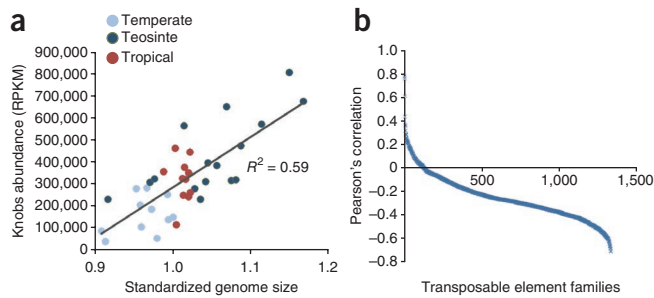


Figure 3 Correlation between knobs, transposable elements and genome size in maize. **(a)** Correlation between knob abundance and genome size. Genome sizes of 17 improved maize lines and 11 teosinte lines correlated with chromosomal knob content (quantified here as reads per kb per million reads mapped (RPKM) of the knob-associated repeats). **(b)** Correlation between abundance of transposable element families and genome size. Most transposable elements were negatively correlated with genome size, and the transposable elements that were significantly positively correlated were associated with chromosomal knobs.

known to contribute to genetic diversity, chromosomal knob segregation rather than global transposable element proliferation is likely to be the major cause of genome size differences within *Z. mays*. In contrast, transposable element abundance seems to explain nearly 50% of the 1.5-fold size difference between maize and *Tripsacum* (Supplementary Tables 15–17). Multiple transposable element families showed higher abundance in *Tripsacum* (Supplementary Table 18), with the remaining variation in genome size likely attributable to *Tripsacum*-specific transposable elements and other repetitive elements not in the maize transposable element database^{6,25}. Taken together, our results support the view that global, genome-wide changes in transposable element content drive genome size difference between grass species^{24,26}, whereas segregation of large, discrete blocks of heterochromatic repeats determine genome size differences within maize.

Knob and transposable element abundance variation, however, do not explain the karyotypic difference between maize and *Tripsacum* ($2n = 2x = 36$). To explore previous suggestions of a shared allopolyploid event in the history of *Zea* and *Tripsacum* lineages^{27,28}, we mapped the proteins encoded by *Tripsacum* reads against those from maize and *Sorghum bicolor*^{7,29} and found that 97.8% of genic reads mapped to maize proteins (Supplementary Table 19). A mere 0.28% of all *Tripsacum* reads showed a closer relationship to *Sorghum* than to maize, effectively ruling out contribution from a non-*Zea* genome. Furthermore, *Tripsacum* reads mapped to the maize reference genome with notably even coverage (Supplementary Fig. 8), suggesting that large-scale structural variations have not occurred since the genera diverged and that observed karyotypic differences are probably the result of chromosome fission. Given that *Tripsacum* has successfully adapted to a wide range of environments (from South America to Iowa), the similarity between the genomes suggests that *Tripsacum* genetics should be investigated for use in the improvement of maize, as mining genetic variation in *Tripsacum* could be very productive.

By providing an unprecedented density of polymorphic markers, the HapMap2 data set we have generated here presents a significant resource for association mapping, genomic selection and the mining of genomic regions that have been selected during domestication and improvement³⁰. In many plant species, highly repetitive genomes and structural diversity complicate access to genetic markers for

germplasm improvement. As maize is not only an economically important crop but also a model for complex genomes, it is anticipated that many of the methods developed here will accelerate genetic variation discovery in other crops.

URLs. Panzea, <http://www.panzea.org/>; Novoalign, <http://www.novocraft.com/>.

METHODS

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

Accession codes. Sequencing data generated in this study have been deposited at the NCBI Short Read Archive under the accession SRA051245. Maize HapMap2 genotypes and other auxiliary data can be found on the Panzea website (see URLs).

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

ACKNOWLEDGMENTS

This work was supported by the US National Science Foundation (DBI-0820619, 0321467, 0703908, 0638566 and IOS-092270), the USDA-ARS, the USDA–National Institute of Food and Agriculture (NIFA) (2009-01864), the US DOE (BER DE-FC02-07ER44494 and DE-AC02-03CH11211), The Rockefeller Foundation, the Bill and Melinda Gates Foundation, the Generation Challenge Program, the Chinese 971 program (2007CB813701, 2007CB813701 and 2007CB813703), the National Natural Science Foundation of China (NSFC) to Young Scientists (10723008), Guangdong Innovation Team Funding, the Chinese Ministry of Agriculture 984 program (2010-Z11), the National High Technology Research and Development Program of China (2009AA10AA03-2) and the National Basic Research Program of China (2007CB108900).

AUTHOR CONTRIBUTIONS

The manuscript was prepared by J.-M.C., B.G., E.S.B., M.D.M., J.R.-I. and D.W. Data analyses (including read mapping, variant detection, scoring and functional analyses) were performed by J.-M.C., C.S., J.C.G., M.G., M.B.H., T.P., Q.S., M.I.T., X.X., J.R.-I. and E.S.B. Transposon mapping and genome size analyses were performed by J.-M.C., M.G., D.C., M.I.T., J.R.-I. and B.G. *Tripsacum* analyses were provided by Q.S., D.C., J.C.G. and E.S.B. GWAS analyses were performed by P.J.B., M.L., F.T. and Z.Z. N.d.L., R.N., J.P., R.S.S. and S.M.K. provided early access data. J.D., R.J.E., L.G., J.C.G., K.E.G., J.H., J.L., X.L., Y.L., R.M., B.M.P., T.R., J.W., S.M.K., X.X., M.D.M., G.Z. and Y.X. provided germplasm management, developed DNA libraries and/or performed sequencing experiments. J.H., J.L., J.W., M.D.M., X.X., E.S.B. and D.W. provided experimental design and coordination.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at <http://www.nature.com/doi/10.1038/ng.2313>.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

- 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).
- Wright, S.I. *et al.* The effects of artificial selection on the maize genome. *Science* **308**, 1310–1314 (2005).
- Fu, H. & Dooner, H. Intraspecific violation of genetic colinearity and its implications in maize. *Proc. Natl. Acad. Sci. USA* **99**, 9573–9578 (2002).
- Morgante, M., de Paoli, E. & Radovic, S. Transposable elements and the plant pan-genomes. *Curr. Opin. Plant Biol.* **10**, 149–155 (2007).
- Swanson-Wagner, R.A. *et al.* Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res.* **20**, 1689–1699 (2010).
- Baucom, R.S. *et al.* Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genet.* **5**, e1000732 (2009).
- Schnable, P.S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
- Le Rouzic, A., Boutin, T.S. & Capy, P. Long-term evolution of transposable elements. *Proc. Natl. Acad. Sci. USA* **104**, 19375–19380 (2007).
- Gore, M.A. *et al.* A first-generation haplotype map of maize. *Science* **326**, 1115–1117 (2009).

10. Brunner, S., Fengler, K., Morgante, M., Tingey, S. & Rafalski, A. Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell* **17**, 343–360 (2005).
11. McMullen, M.D. *et al.* Genetic properties of the maize nested association mapping population. *Science* **325**, 737–740 (2009).
12. Li, R. *et al.* SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19**, 1124–1132 (2009).
13. Li, H. & Homer, N. A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform.* **11**, 473–483 (2010).
14. Clark, R.M. *et al.* Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* **317**, 338–342 (2007).
15. Rizzon, C., Ponger, L. & Gaut, B.S. Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. *PLoS Comput. Biol.* **2**, e115 (2006).
16. Lai, J. *et al.* Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat. Genet.* **42**, 1027–1030 (2010).
17. 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).
18. 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).
19. Tian, F. *et al.* Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* **43**, 159–162 (2011).
20. SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. & Bennetzen, J.L. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**, 43–45 (1998).
21. Albert, P.S., Gao, Z., Danilova, T.V. & Birchler, J.A. Diversity of chromosomal karyotypes in maize and its relatives. *Cytogenet. Genome Res.* **129**, 6–16 (2010).
22. Laurie, D. & Bennett, M. Nuclear DNA content in the genera *Zea* and *Sorghum*. Intergeneric, interspecific and intraspecific variation. *Heredity* **55**, 307–313 (1985).
23. Poggio, L., Rosato, M. & Chiavarino, A. Genome size and environmental correlations in maize (*Zea mays* ssp. *mays*, Poaceae). *Ann. Bot.* **82**, 107–115 (1998).
24. Tenailon, M.I., Hufford, M.B., Gaut, B.S. & Ross-Ibarra, J. Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. *Genome Biol. Evol.* **3**, 219–229 (2011).
25. Meyers, B.C., Tingey, S.V. & Morgante, M. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res.* **11**, 1660–1676 (2001).
26. Piegue, B. *et al.* Doubling genome size without polyploidization: dynamics of retrotransposon-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* **16**, 1262–1269 (2006).
27. Anderson, E. Cytological observations on *Tripsacum dactyloides*. *Ann. Mo. Bot. Gard.* **31**, 317–323 (1944).
28. Gaut, B.S., Le Thierry Ennequin, M., Peek, A.S. & Sawkins, M.C. Maize as a model for the evolution of plant nuclear genomes. *Proc. Natl. Acad. Sci. USA* **97**, 7008–7015 (2000).
29. Paterson, A.H. *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556 (2009).
30. Hufford, M.B. *et al.* Comparative population genomics of maize domestication and improvement. *Nat. Genet.* published online, doi:10.1038/ng.2309 (3 June 2012).

ONLINE METHODS

Samples and Illumina sequencing. Total genomic DNA was isolated from homogenized frozen ear tissue, lyophilized leaf tissue or etiolated seedling material of 103 inbred lines using the standard cetyltrimethyl ammonium bromide (CTAB) DNA extraction protocol. Covaris-sheared total genomic DNA was prepared following the standard Illumina paired-end library protocol, ultimately capturing fragments in the 250-nt range, and delivered for paired-end sequencing of 76–100 bp on the Genome Analyzer IIX platform (Illumina).

Read mapping, variant identification and scoring. Two complementary pipelines were used to map the reads and identify variant sites, and a scoring model was developed to filter for high-quality polymorphisms resulting from both pipelines. Reads were mapped against the B73 reference sequence⁷ (version AGPv1). The first pipeline, which was previously used to align unpaired short reads from HapMap1 (ref. 9), uses Bowtie³¹ and Novoalign for mapping reads. Only uniquely mapping reads were retained. The SAMtools package³² was used to filter for PCR duplicates and identify putative variation sites. The second pipeline use SOAP2 for read mapping³³, realSFS³⁴ for identifying putative variation sites and SOAPsnp¹² for calling the genotype of each individual. At this stage, variations were filtered for alternate allele quality and homozygosity (Phred score of ≥ 20 , with $\geq 90\%$ of the lines homozygous and with heterozygous genotypes being no more than double the frequency of the homozygous minor allele genotypes).

A logistic regression model was used to score the SNPs. The factors in the model included the proportion of homozygous lines, a segregation test and an LD test with an anchor map. This anchor map was a small subset of the sites with (i) a presence in at least 50 lines, (ii) homozygosity in 90% of the lines, (iii) an alternate allele with average quality greater than 20, (iv) minor allele with homozygosity in four lines, (v) contingency test $-\log_{10}(P) > 2.5$, (vi) LD test $-\log_{10}(P) > 5$ and (vii) the imputation of the minor allele correct more than 80% of the time (a test of haplotype structure). The segregation test is a contingency test of read depth for SNP allele by line, with significance determined by permutations⁹.

Regions of IBD were used to build and test the scoring model, with the expectation that there should be no SNP differences within pairs of IBD segments. The inbred lines were previously genotyped with an Illumina MaizeSNP50 array (M.D.M. and J.R.-I., unpublished data), and IBD regions were identified from these data by identifying unbroken stretches of at least 150 identical SNP alleles between pairs of lines. The distribution of IBD blocks across the genome, in relation to recombination rate, is presented in **Supplementary Figure 9**. The model coefficients were then determined using the 55 pairs of IBD blocks localized to the pericentromeric region along chromosome 10 (between positions 47–56 Mb in AGPv1), including only sites that could be tested in at least 20 pairs of contrasts.

After filtering and scoring, 55,061,920 variants remained, including 3.2 million indels. The size distribution of indels is shown in **Supplementary Figure 10**. The coverage of variations across the whole genome is plotted for each chromosome (**Supplementary Fig. 11**). These scored variants agree very well with previous genotyping results, with discordant rates ranging from 0.10–1.57% (**Supplementary Tables 20–23** and **Supplementary Note**).

LD decay. LD (r^2)³⁵ was estimated for all pairs of sites within 300 kb of each other that were homozygous for at least two minor alleles and present in at least 40 lines (minimum allele frequency of 5%). We also conducted a test to evaluate how close we were to complete genome coverage in LD. We did this by calculating the maximum LD for all SNPs within a 300-kb window (using the above position and minor allele cutoffs) (**Supplementary Table 24**).

Read-depth variation. RDVs were identified for each inbred line by first counting the number of reads mapping to nonoverlapping sliding 10-kb windows across the B73 reference assembly (10-kb RDVs) as well as within genic loci (gene RDVs). Genic loci were defined by the 32,450 filtered gene set annotated on the B73 reference genome (release 4a.53), with an additional 2 kb included at both the 5' and 3' ends of the genes. Only mapping results from the paired-end libraries using the Bowtie-*Novoalign* pipeline were used in this analysis.

The read-depth counts of each line were then compared against a high-coverage B73 sequence library. This B73 library consists of 362 million 76-bp paired-end reads, giving $\sim 25\times$ coverage of the genome (C.S., X.X. and G.Z., unpublished data). EdgeR³⁶, a Bioconductor³⁷ package for analyzing digital gene expression, was then adapted to estimate \log_2 coverage ratios of each line against this high-depth B73 library (M. Robinson, personal communication). A 10-kb window or gene was considered as having a significant RDV if there was a twofold change compared to the high-coverage B73 library with a false discovery rate of ≤ 0.01 .

The 10% most variant and 10% least variant genes were identified and are listed in **Supplementary Table 25**. The *Arabidopsis* and rice orthologs for the most RDV-variant genes are listed in **Supplementary Table 26**.

RDV anchoring. To determine whether RDVs were anchored locally, we used LD by implementing a simple *t* test between the 10-kb interval of an RDV (a quantitative character) and the SNP genotype. Key to this contrast is that we looked for LD between 'missing' regions of the genome with SNPs that were 'present', which prevents extended anomalies from appearing as LD. We only used SNPs that were present in more than 70% of the lines and that had the minor allele in more than five lines (minimum minor allele frequency of $>5\%$). All SNPs within 100 kb of the RDV interval were tested. A simple Bonferroni correction was applied to control for the difference in the number of SNP tests for each interval. To investigate the importance of including SNPs from within the RDV interval, in one version of the test, we excluded all SNPs from within the interval. This yielded only a very minor difference in the results. Because population structure can produce significant results, even for unlinked RDVs, we also conducted randomization tests in which we evaluated the distribution of *P* values, but only for sites that were on the same chromosome but over 500,000 bp away. This distribution identified *P* values that were likely the result of population structure from the sites that were almost likely the product of local LD.

GWAS in NAM. GWAS on the NAM population for five traits using the combined HapMap1 and HapMap2 data sets was conducted using a previously described method¹⁹, with BPP—defined as the proportion of times a SNP is included in the model—used to evaluate the strength of detected associations. Associations with BPP of 0.05 or greater were used for further analysis.

Assessment of transposable element and knob content. In order to assess transposable element and knob repeat content across the 103 maize and teosinte inbred lines, we followed the SSAHA2 (ref. 38) mapping protocol described in a recent publication²⁴, with the exception that additional comparisons were performed against 180-bp and 360-bp knob-specific tandem repeats (GenBank, M32522.1 and AF071124.1, respectively). The estimates of knob abundance in each of the 103 lines are listed in **Supplementary Table 27**.

Flow cytometry. The protocol for the preparation of leaf samples for flow cytometry used in this study is based on a previously described protocol³⁹, with slight modifications (**Supplementary Note**). As the genome sizes were estimated in two separate experiments, ANOVA models were fitted separately for each experiment with PROC MIXED in SAS statistical software (SAS Institute). The fitted models had genome size as the dependent variable, line as a fixed effect and rep nested within line as a random effect. Degrees of freedom were calculated via the Satterthwaite approximation. Least-square means were obtained with the LSMEANS statement in PROC MIXED. Then, these means were standardized to the mean genome size of the B73 inbred reference standard included in each experiment. This was necessary to permit joint analysis of the genome size data from both experiments.

Interspecies comparison of gene content. *Tripsacum* reads were matched using BlastX⁴⁰ against maize⁷ and *Sorghum*²⁹ proteins, with *S. bicolor* chosen to represent a non-*Zea* grass genome. Reads that corresponded to either maize or sorghum proteins were then mapped against the sorghum (v1 assembly) and maize (RefGenV2) reference genomes using BlastN to identify reads that had closer homology to sorghum than maize.

31. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
32. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
33. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).
34. Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010).
35. Hill, W.G. & Robertson, A. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38**, 226–231 (1968).
36. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
37. Gentleman, R.C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
38. Ning, Z., Cox, A.J. & Mullikin, J.C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).
39. Arumuganathan, K. & Earle, E.D. Estimation of nuclear DNA content of plants by flow cytometry. *Plant Mol. Biol. Rep.* **9**, 229–241 (1991).
40. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).