A First-Generation Haplotype Map of Maize
Michael A. Gore, *et al.*
Science **326**, 1115 (2009);
DOI: 10.1126/science.1177837

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

11. F. Wei *et al.*, *PLoS Genet.*, 19 November 2009 (10.1371/journal.pgen.1000715).
12. F. Wei *et al.*, *PLoS Genet.* **3**, e123 (2007).
13. S. Zhou *et al.*, *PLoS Genet.*, 19 November 2009 (10.1371/journal.pgen.1000711).
14. Materials and methods are available as supporting material on *Science* Online.
15. P. SanMiguel *et al.*, *Science* **274**, 765 (1996).
16. B. McClintock, *Cold Spring Harbor Symp. Quant. Biol.* **16**, 13 (1951).
17. C. Feschotte, N. Jiang, S. R. Wessler, *Nat. Rev. Genet.* **3**, 329 (2002).
18. A. Kumar, J. L. Bennetzen, *Annu. Rev. Genet.* **33**, 479 (1999).
19. S. Liu *et al.*, *PLoS Genet.*, 19 November 2009 (10.1371/journal.pgen.1000733).
20. V. V. Kapitonov, J. Jurka, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 8714 (2001).
21. S. Lal, N. Georgelis, L. Hannah, in *Handbook of Maize: Genetics and Genomics*, J. L. Bennetzen, S. Hake, Eds. (Springer, New York, 2008), pp. 329–339.
22. L. Yang, J. L. Bennetzen, *Proc. Natl. Acad. Sci. USA*, published online 19 November 2009 (10.1073/pnas.0908008106).
23. L. Yang, J. L. Bennetzen, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 12832 (2009).
24. R. S. Baucom *et al.*, *PLoS Genet.*, 19 November 2009 (10.1371/journal.pgen.1000732).
25. F. Wei *et al.*, *PLoS Genet.*, 19 November 2009 (10.1371/journal.pgen.1000728).
26. L. Zhang, *PLoS Genet.*, 19 November 2009 (10.1371/journal.pgen.1000716).
27. H. Liang, W. H. Li, *Mol. Biol. Evol.* **26**, 1195 (2009).
28. G. Haberer *et al.*, *Plant Physiol.* **139**, 1612 (2005).
29. N. N. Alexandrov *et al.*, *Plant Mol. Biol.* **69**, 179 (2009).
30. C. Soderlund *et al.*, *PLoS Genet.*, 19 November 2009 (10.1371/journal.pgen.1000740).
31. T. K. Wolfgruber *et al.*, *PLoS Genet.*, 19 November 2009 (10.1371/journal.pgen.1000743).
32. C. X. Zhong *et al.*, *Plant Cell* **14**, 2825 (2002).
33. A. Sharma, G. G. Presting, *Mol. Genet. Genomics* **279**, 133 (2008).
34. A. Sharma, K. L. Schneider, G. G. Presting, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 15470 (2008).
35. D. Lisch, *Annu. Rev. Plant Biol.* **60**, 43 (2009).
36. M. Alleman *et al.*, *Nature* **442**, 295 (2006).
37. Y. Jia *et al.*, *PLoS Genet.*, 19 November 2009 (10.1371/journal.pgen.1000737).
38. Y. Fu *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 12282 (2005).
39. L. E. Palmer *et al.*, *Science* **302**, 2115 (2003).
40. W. Zhang, H. R. Lee, D. H. Koo, J. Jiang, *Plant Cell* **20**, 25 (2008).
41. M. A. Gore *et al.*, *Science*, **326**, 1115 (2009).
42. R. A. Swanson-Wagner *et al.*, *Science* **326**, 1118 (2009).
43. N. M. Springer *et al.*, *PLoS Genet.*, 19 November 2009 (10.1371/journal.pgen.1000734).
44. C. G. Tian *et al.*, *Yi Chuan Xue Bao* **32**, 519 (2005).
45. C. Seoighe, C. Gehring, *Trends Genet.* **20**, 461 (2004).
46. B. W. Penning *et al.*, *Plant Physiol.*, published online 19 November 2009 (10.1104/pp.109.136804).
47. H. Shaked, K. Kashkush, H. Ozkan, M. Feldman, A. A. Levy, *Plant Cell* **13**, 1749 (2001).
48. K. Song, P. Lu, K. Tang, T. C. Osborn, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 7719 (1995).
49. J. A. Tate, P. Joshi, K. A. Soltis, P. S. Soltis, D. E. Soltis, *BMC Plant Biol.* **9**, 80 (2009).
50. B. C. Thomas, B. Pedersen, M. Freeling, *Genome Res.* **16**, 934 (2006).
51. S. J. Emrich *et al.*, *Genetics* **175**, 429 (2007).
52. B. McClintock, *Science* **69**, 629 (1929).
53. The Maize Genome Sequencing Project supported by NSF award DBI-0527192 (R.K.W., S.W.C., R.S.F., R.A.W., P.S.S., S.A., L.S., D.W., W.R.M., R.A.M.). The Maize Transposable Element Consortium and the Maize Centromere Consortium supported by NSF awards DBI-0607123 (S.R.W., J.L.B., R.K.D., N.J., P.S.M.) and DBI-0421671 (R.K.D., J.J., G.G.P.). Also supported by NSF grants DBI-0321467 (D.W.), DBI-0321711 (P.S.S.), DBI-0333074 (D.W.), DBI-0501818 (D.C.S.), DBI-0501857 (Y.Y.), DBI-0701736 (T.P.B., Q.S.), DBI-0703273 (R.A.M.), and DBI-0703908 (D.W.), and by USDA National Research Initiative Grants 2005-35301-15715 and 2007-35301-18372 from the USDA Cooperative State Research, Education, and Extension Service (P.S.S.) and from the USDA-ARS (408934 and 413089) to D.W., and from the Office of Science (Biological and Environmental Research), U.S. Department of Energy, grant DE-FG02-08ER64702 to N.C.C. and M.C.M. Sequences of the reference chromosomes have been deposited in GenBank as accession numbers CM000777 to CM000786. RNA-sequence reads have been deposited in the Gene Expression Omnibus (GEO) database (www.ncbi.nlm.nih.gov/geo) as accession numbers GSE16136, GSE16868, and GSE16916. Centromeric sequences have been deposited in the National Center for Biotechnology Information, NIH, Trace Archive as accessions 1757396377 to 1757412600 and 2185189231 to 2185200942.

Supporting Online Material

www.sciencemag.org/cgi/content/full/326/5956/1112/DC1
Materials and Methods
SOM Text
Figs. S1 to S18
Tables S1 to S18
References

1 July 2009; accepted 13 October 2009
10.1126/science.1178534

A First-Generation Haplotype Map of Maize

Michael A. Gore,[1,2,3]*† Jer-Ming Chia,[4]* Robert J. Elshire,[3] Qi Sun,[5] Elhan S. Ersoz,[3] Bonnie L. Hurwitz,[4]‡ Jason A. Peiffer,[2] Michael D. McMullen,[1,6] George S. Grills,[7] Jeffrey Ross-Ibarra,[8] Doreen H. Ware,[1,4]§ Edward S. Buckler[1,2,3]§

Maize is an important crop species of high genetic diversity. We identified and genotyped several million sequence polymorphisms among 27 diverse maize inbred lines and discovered that the genome was characterized by highly divergent haplotypes and showed 10- to 30-fold variation in recombination rates. Most chromosomes have pericentromeric regions with highly suppressed recombination that appear to have influenced the effectiveness of selection during maize inbred development and may be a major component of heterosis. We found hundreds of selective sweeps and highly differentiated regions that probably contain loci that are key to geographic adaptation. This survey of genetic diversity provides a foundation for uniting breeding efforts across the world and for dissecting complex traits through genome-wide association studies.

Maize (*Zea mays* L.) is both a model genetic system and an important crop species. Already a critical source of food, fuel, feed, and fiber, the addition of genomic information allows maize to be further improved through plant breeding that exploits its tremendous genetic diversity (*1*–*3*). Genome-wide association studies (GWAS) of diverse maize germplasm offer the potential to rapidly resolve complex traits to gene-level resolution, but these studies require a high density of genome-wide markers. To do this, we targeted the 20% of the maize genome that is low-copy (*4*, *5*) on a diverse panel of 27 inbred lines (representative of maize breeding efforts and worldwide diversity)—founders of the maize nested association mapping (NAM) population (*6*)—and used sequencing-by-synthesis (SBS) technology with three complementary restriction enzyme–anchored genomic libraries (figs. S1 and S2A) (*7*).

More than 1 billion SBS reads (>32 gigabases of sequence) were generated, covering ~38% of the total maize genome, albeit at mostly low-coverage levels. We focused on the ~93 million base pairs (Mbp) of low-copy sequence present in 13 or more lines in this study. Roughly 39% of the sequenced low-copy fraction was derived from introns and exons (*5*), covering 32% of the total genic fraction in the genome. We identified 3.3 million single-nucleotide polymorphisms (SNPs) and indels (table S1) and found that, overall, 1 in every 44 bp was polymorphic ($\pi = 0.0066$ per base pair). In a subset used for the population genetics analyses, the error rate was 1/2570 or 17-fold lower than π (roughly half the errors are paralogy issues). The absolute level of diversity we examined, though high, may be slightly reduced because of difficulties aligning highly divergent sequences and our low power to call

[1]United States Department of Agriculture–Agriculture Research Service (USDA-ARS). [2]Department of Plant Breeding and Genetics, Cornell University, Ithaca, NY 14853, USA. [3]Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853, USA. [4]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA. [5]Computational Biology Service Unit, Cornell University, Ithaca, NY 14853, USA. [6]Division of Plant Sciences, University of Missouri, Columbia, MO 65211, USA. [7]Institute for Biotechnology and Life Science Technologies, Cornell University, Ithaca, NY 14853, USA. [8]Department of Plant Sciences, University of California, Davis, CA 95616–5294, USA.

*These authors contributed equally to this work.
†Present address: United States Arid-Land Agricultural Research Center, Maricopa, AZ 85138, USA.
‡Present address: Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA.
§To whom correspondence should be addressed. E-mail: ware@cshl.edu (D.H.W.); esb33@cornell.edu (E.S.B.)

singleton variants [60% of the expected rate on the basis of Sanger sequencing of candidate gene amplicons (*8*)].

Due to duplications of an ancestral genome, maize has many paralogous regions (*5*), and, as a result, 41% of the identified polymorphisms appear to be differences between paralogous sequences in these inbred lines. We thus defined two sets of SNPs: (i) the association set and (ii) the diversity set. Paralogous variants can be used effectively for GWAS and were retained in our association SNP data set, but they pose problems for analyses of diversity and were removed from our diversity SNP data set. The diversity set provides SNPs for characterizing genome-wide variation patterns, whereas the association set provides access to more regions of the genome. Comparisons between pairs of maize inbred lines identified structural variation for both retrotransposons and gene fragments (*9*, *10*). Similarly, our diverse lines averaged an excess of 7.8% of reads that were unique or unalignable to the reference genome (fig. S2B). On the basis of these data, the B73 genome may only capture ~70% of the alignable low-copy fraction represented by these 27 lines. Capturing the entire genome space for maize will be critical to evaluation of the functional importance of such divergent sequences.

In spite of the considerable molecular variation in the maize genome, the evolutionary potential of many variants is limited by linkage. Because this HapMap was built on the 27 founders of the NAM population, which captures ~135,000 meiotic crossovers (*6*), we could compare estimates of the recombination rate (*R*) with historical recombination patterns inferred on the basis of the SNP distribution (ρ). Overall, *R* and ρ were strongly correlated, indicating that recombination patterns tend to be stable over time [Spearman correlation r^2_{sp} = 0.56 (Fig. 1B and fig. S3B)]. At the chromosomal scale, total genic bases were nearly perfectly correlated (probably the euchromatin fraction) with the total *R* on the basis of the NAM [r^2_{sp} = 0.88 (Fig. 1A and fig. S3A)]. Recombination varied dramatically along

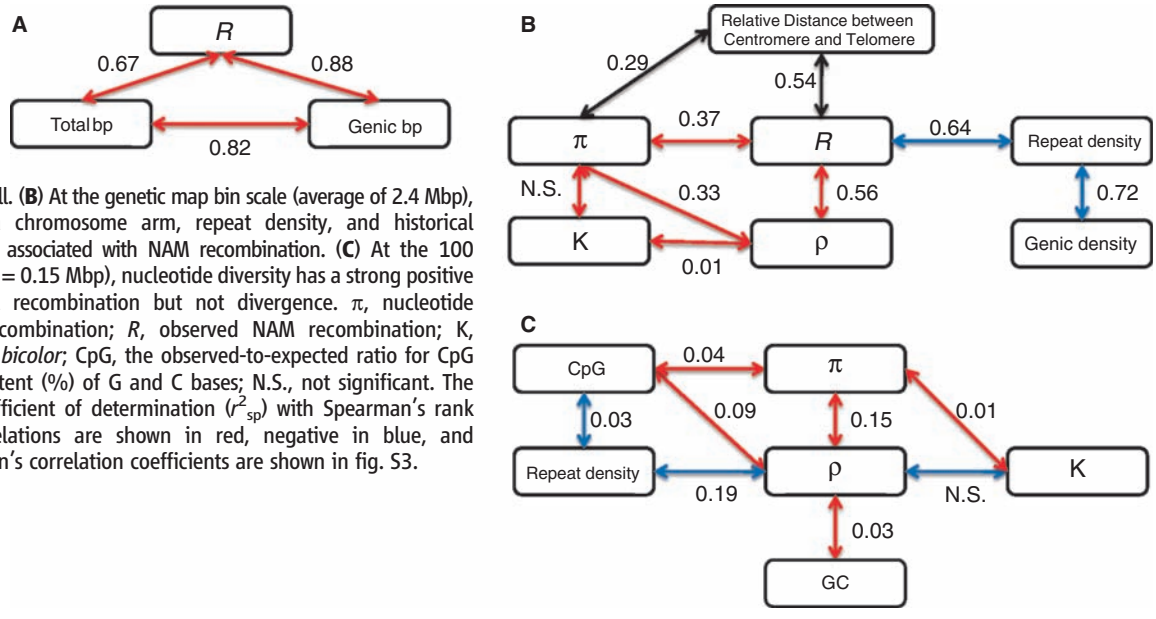

Fig. 1. Relation between sequence features, recombination, and diversity at three scales. (**A**) At the chromosomal scale (average of 200 Mbp), the total genic size of a chromosome predicts total recombination well. (**B**) At the genetic map bin scale (average of 2.4 Mbp), relative distance along a chromosome arm, repeat density, and historical recombination are strongly associated with NAM recombination. (**C**) At the 100 SNP bin scale (average size = 0.15 Mbp), nucleotide diversity has a strong positive correlation with historical recombination but not divergence. π, nucleotide diversity; ρ, historical recombination; *R*, observed NAM recombination; K, divergence from *Sorghum bicolor*; CpG, the observed-to-expected ratio for CpG dinucleotides; GC, the content (%) of G and C bases; N.S., not significant. The numbers indicate the coefficient of determination (r^2_{sp}) with Spearman's rank correlation. Positive correlations are shown in red, negative in blue, and symmetric in black. Pearson's correlation coefficients are shown in fig. S3.

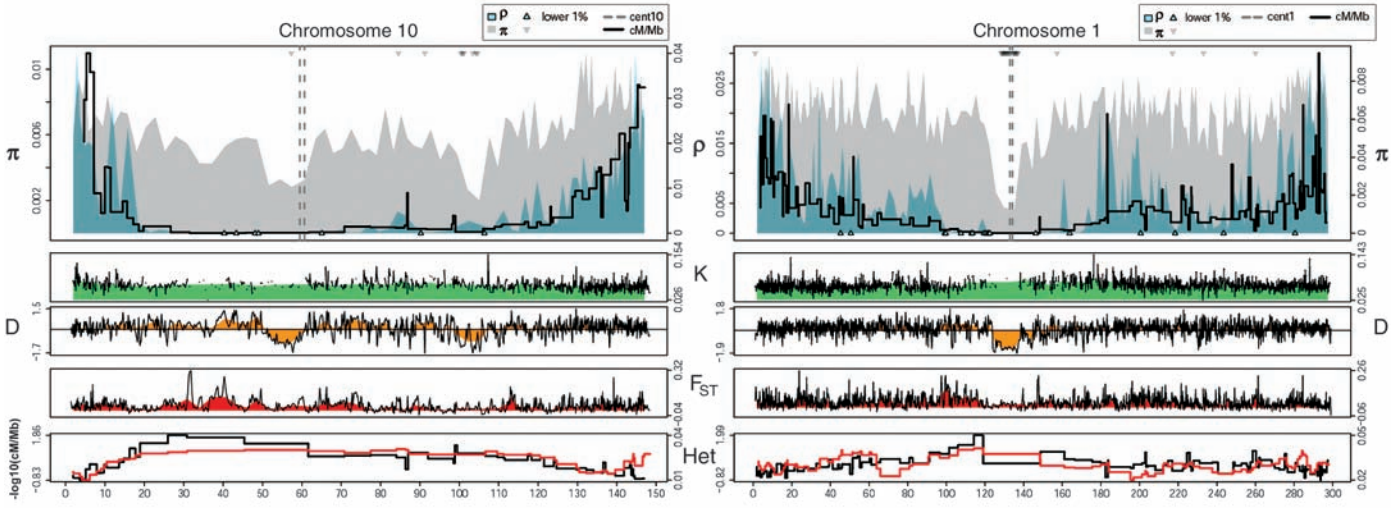

Fig. 2. Diversity along maize chromosomes 1 and 10. The horizontal axes are in units of million base pairs along the B73 reference genome; centromeres (*25*) are delineated by vertical dotted lines. Each chromosome shows (top panel) nucleotide diversity (π) and historical (ρ) and observed (cM/Mb, *R*) recombination (bottom panels) divergence from *Sorghum bicolor* (K), Tajima's measure of the site-frequency spectrum (D), population differentiation (F_{ST}), and a comparison of recombination [-log10(cM/Mb)] to residual heterozygosity (Het). Filled polygons represent the median of population genetic data from 10 100-SNP windows. Thin black lines denote data plotted for individual windows; thicker lines indicate data for residual heterozygosity and recombination represent estimates over NAM genetic map bins.

the chromosome (Fig. 2 and fig. S4), with 95% of total *R* limited to slightly more than half of the genome. The 90th versus 10th percentiles varied 28-fold for *R* and 12-fold for ρ. All chromosomes had a pericentromeric region of 60 to 113 Mbp with low recombination; these regions contain 21% of the total genic fraction. Similarly, sorghum has large pericentromeric regions that are recombinationally suppressed (*11*), but with fewer genes contained in these regions. We identified two correlated drivers of recombination: (i) the relative distance along a chromosome arm [r^2_{sp} = 0.54 (Fig. 1B and fig. S3B)] and (ii) repeat density [r^2_{sp} = 0.64 (Fig. 1B and fig. S3B)].

An earlier study on the NAM population identified considerable residual heterozygosity in pericentromeric regions of the maize genome and posited that this retention was probably a consequence of heterosis (*6*). We extended this finding by evaluating the relation of residual heterozygosity with recombination rates, genetic variation, and gene density. We found that regions of increased residual heterozygosity ($P < 0.01$) had 36% of all genes and nearly average diversity (91% of the genome average π). By anchoring recombination to the physical genome and controlling for a chromosome effect, residual heterozygosity and *R* were inversely related [$r^2_{sp} = 0.35$ (Fig. 2)], whereas gene density ($r^2_{sp} = 0.18$) and diversity (π) ($r^2_{sp} = 0.16$) were less related. When we control for recombination, gene density and π have a statistically nonsignificant effect on residual heterozygosity. This result indicates that recombination is the major factor determining residual heterozygosity. This result indicates that a relatively low recombination rate is the major factor that contributes to the retention of residual heterozygosity. As a consequence, the tremendous genetic diversity at pericentromeric regions is constrained from being recombined into the most vigorous allelic combinations, thus lending further credence to pseudo-overdominance as the genetic basis for heterotic phenotypes in F_1 hybrids.

Examining nucleotide diversity, we found that chromosomes were punctuated by numerous million base pair–scale valleys of low nucleotide diversity and an excess of low-frequency variants (Fig. 2 and fig. S4). Most notably, 9 of the 10 maize centromeres are in or near such valleys. This observation is consistent with selection and rapid evolution at centromeric regions (*12*) and similar to observations in humans (*13*) and *Drosophila* (*14*), but contrasts with the high pericentromeric diversity in *Arabidopsis* (*15*, *16*). Although most regions of low diversity were associated with centromeres, a large number of low-diversity regions occur throughout the genome, many in regions with considerable recombination.

Genome-wide, nucleotide diversity was correlated with both ρ [$r^2_{sp} = 0.33$ (Fig. 1B)] and *R* [$r^2_{sp} = 0.37$ (Fig. 1B)], but was nearly independent of divergence from *Sorghum* [$r^2_{sp} \leq 0.01$ (Fig. 1, B and C)], indicating that regions of reduced diversity have been the targets of selection. We tested 18 regions that have undergone a selective sweep (table S2), resulting in a median sweep in the 3.1% low tail of nucleotide diversity, suggesting that our HapMap has reasonable power to detect selected regions. In the high-recombination fraction of the genome, we identified 148 regions showing less diversity then the domestication gene *tb1* (*17*): 37 in high-recombination regions and 111 in low-recombination regions, including 1 of 11 megabases in size. A large region identified on the long arm of chromosome 10 has recently been associated with selection during domestication (*18*).

Given the recent divergence of lineages in *Zea* (*19*), selective sweeps may not be associated with domestication, but instead reflect selection in its ancestor, teosinte. Distinguishing between these possibilities and identifying their timing require sampling of diversity in both teosinte and early domesticated varieties (*2*, *20*). Additionally, demographic change has probably contributed to the observed variance in diversity (*2*, *21*), making it difficult to quantify what fraction of low-diversity regions may be due to neutral processes. Hence, investigation of the function and adaptive importance of regions defined here will be an important avenue of future research.

Maize has spread from the tropics into the northern and southern temperate zones and can clearly be differentiated with HapMap SNPs (fig. S5). However, F_{ST} (a statistic that provides a measure of the extent of genetic differentiation between populations) had an average of only 3.8% between temperate and tropical germplasm, which suggests minimal differentiation. Although 43% of the genome has some F_{ST} differentiation ($P < 0.05$), 183 regions showed a highly significant F_{ST} ($P < 0.0001$), and may contain loci involved in the adaptation of maize to temperate versus tropical environments.

GWAS studies require markers in high LD with polymorphisms throughout the genome. This has been challenging, as in diverse maize LD generally decays ($r2 < 0.1$) within 2000 bp (fig. S6) (*22*). However, we also found evidence of longer haplotypes extending for thousands and millions of bases. Association studies in a genome with numerous small QTL effects (*23*) require high LD ($r^2 > 0.8$). We used a SNP hiding test (*24*), which revealed high LD ($r^2 > 0.8$) 55% of the time. When we conducted the same test on SNPs separated by at least 500 bp, high LD was found only 34% of the time. Thus, complete coverage for GWAS may require another order of magnitude of markers and the ability to anchor markers into the middle of retrotransposon domains.

With the maize HapMap and genome, we identified evidence for hundreds of regions that are probably involved in domestication and the geographic differentiation of maize. Remarkably, all of this selection has had to work against a genome with very strong recombinational suppression, which has effects that are embodied in modern-day heterosis and ancient, massive sweeps in centromeric regions. The future of maize improvement will not only depend on the ability to identify favorable alleles from the world's germplasm, but also the application of selection in a manner that effectively overcomes these recombinational constraints.

References and Notes

1. M. I. Tenaillon *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 9161 (2001).
2. S. I. Wright *et al.*, *Science* **308**, 1310 (2005).
3. S. A. Flint-Garcia *et al.*, *Plant J.* **44**, 1054 (2005).
4. P. SanMiguel *et al.*, *Science* **274**, 765 (1996).
5. P. S. Schnable *et al.*, *Science* **326**, 1112 (2009).
6. M. D. McMullen *et al.*, *Science* **325**, 737 (2009).
7. Materials and methods are available as supporting material on *Science* Online.
8. www.panzea.org
9. H. Fu, H. K. Dooner, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 9573 (2002).
10. M. Morgante *et al.*, *Nat. Genet.* **37**, 997 (2005).
11. A. Paterson *et al.*, *Nature* **457**, 551 (2009).
12. S. Henikoff, K. Ahmad, H. S. Malik, *Science* **293**, 1098 (2001).
13. I. Hellmann *et al.*, *Genome Res.* **18**, 1020 (2008).
14. D. J. Begun *et al.*, *PLoS Biol.* **5**, e310 (2007).
15. R. M. Clark *et al.*, *Science* **317**, 338 (2007).
16. A. Kawabe, A. Forrest, S. I. Wright, D. Charlesworth, *Genetics* **179**, 985 (2008).
17. R.-L. Wang, A. Stec, J. Hey, L. Lukens, J. Doebley, *Nature* **398**, 236 (1999).
18. F. Tian, N. M. Stevens, E. S. Buckler, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 9979 (2009).
19. J. Ross-Ibarra, M. Tenaillon, B. S. Gaut, *Genetics* **181**, 1399 (2009).
20. M. Yamasaki *et al.*, *Plant Cell* **17**, 2859 (2005).
21. K. R. Thornton, J. D. Jensen, C. Becquet, P. Andolfatto, *Heredity* **98**, 340 (2007).
22. D. L. Remington *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 11479 (2001).
23. E. S. Buckler *et al.*, *Science* **325**, 714 (2009).
24. S. Kim *et al.*, *Nat. Genet.* **39**, 1151 (2007).
25. T. K. Wolfgruber *et al.*, *PLoS Genet.* **5**, e1000743 (2009).

26. We thank D. Costich and L. Rigamer Lirette for technical editing of the manuscript; researchers at the Lita Annenberg Hazen Genome Sequencing Center of Cold Spring Harbor Laboratory for discussion about sequencing and library construction; and T. Stelick, P. Schweitzer, and J. I. VanEe for assistance with the SBS data, all of which was generated at the Cornell University Life Sciences Core Laboratories Center. Mention of trade names or commercial products was solely to provide specific information and does not imply recommendation or endorsement by the USDA. This work was supported by NSF grants DBI-0321467, DBI-0638566, and DBI-0820619, and by the USDA-ARS. Sequences have been deposited at National Center for Biotechnology Information Short Read Archive with accession number SRP001145, and SNP calls are available at www.panzea.org.

Supporting Online Material

www.sciencemag.org/cgi/content/full/326/5956/1115/DC1
Materials and Methods
Figs. S1 to S7
Tables S1 and S2
References

17 June 2009; accepted 20 October 2009
10.1126/science.1177837