

The B73 Maize Genome: Complexity, Diversity, and Dynamics

Patrick S. Schnable,^{1,2,3,4*} Doreen Ware,^{5,6*} Robert S. Fulton,^{7†} Joshua C. Stein,^{6†} Fusheng Wei,^{8†} Shiran Pasternak,⁶ Chengzhi Liang,⁶ Jianwei Zhang,⁸ Lucinda Fulton,⁷ Tina A. Graves,⁷ Patrick Minx,⁷ Amy Denise Reily,⁷ Laura Courtney,⁷ Scott S. Kruchowski,⁷ Chad Tomlinson,⁷ Cindy Strong,⁷ Kim Delehaunty,⁷ Catrina Fronick,⁷ Bill Courtney,⁷ Susan M. Rock,⁷ Eddie Belter,⁷ Feiyu Du,⁷ Kyung Kim,⁷ Rachel M. Abbott,⁷ Marc Cotton,⁷ Andy Levy,⁷ Pamela Marchetto,⁷ Kerri Ochoa,⁷ Stephanie M. Jackson,⁷ Barbara Gillam,⁷ Weizu Chen,⁷ Le Yan,⁷ Jamey Higginbotham,⁷ Marco Cardenas,⁷ Jason Waligorski,⁷ Elizabeth Applebaum,⁷ Lindsey Phelps,⁷ Jason Falcone,⁷ Krishna Kanchi,⁷ Thynn Thane,⁷ Adam Scimone,⁷ Nay Thane,⁷ Jessica Henke,⁷ Tom Wang,⁷ Jessica Ruppert,⁷ Neha Shah,⁷ Kelsi Rotter,⁷ Jennifer Hodges,⁷ Elizabeth Ingenthron,⁷ Matt Cordes,⁷ Sara Kohlberg,⁷ Jennifer Sgro,⁷ Brandon Delgado,⁷ Kelly Mead,⁷ Asif Chinwalla,⁷ Shawn Leonard,⁷ Kevin Crouse,⁷ Kristi Collura,⁸ Dave Kudrna,⁸ Jennifer Currie,⁸ Ruifeng He,⁸ Angelina Angelova,⁸ Shanmugam Rajasekar,⁸ Teri Mueller,⁸ Rene Lomeli,⁸ Gabriel Scara,⁸ Ara Ko,⁸ Krista Delaney,⁸ Marina Wissotski,⁸ Georgina Lopez,⁸ David Campos,⁸ Michele Braidotti,⁸ Elizabeth Ashley,⁸ Wolfgang Golser,⁸ HyeRan Kim,⁸ Seunghye Lee,⁸ Jinke Lin,⁸ Zeljko Dujmic,⁸ Woojin Kim,⁸ Jayson Talag,⁸ Andrea Zuccolo,⁸ Chuanzhu Fan,⁸ Aswathy Sebastian,⁸ Melissa Kramer,⁶ Lori Spiegel,⁶ Lidia Nascimento,⁶ Theresa Zutavern,⁶ Beth Miller,⁶ Claude Ambroise,⁶ Stephanie Muller,⁶ Will Spooner,⁶ Apurva Narechania,⁶ Liya Ren,⁶ Sharon Wei,⁶ Sunita Kumari,⁶ Ben Faga,⁶ Michael J. Levy,⁶ Linda McMahan,⁶ Peter Van Buren,⁶ Matthew W. Vaughn,⁶ Kai Ying,³ Cheng-Ting Yeh,^{1,2} Scott J. Emrich,^{9,10} Yi Jia,³ Ananth Kalyanaraman,^{9,11} An-Ping Hsia,^{1,2} W. Brad Barbazuk,¹² Regina S. Baucum,¹³ Thomas P. Brutnell,¹⁴ Nicholas C. Carpita,¹⁵ Cristian Chaparro,¹⁶ Jer-Ming Chia,⁶ Jean-Marc Deragon,¹⁶ James C. Estill,^{13,17} Yan Fu,^{2,4} Jeffrey A. Jeddelloh,¹⁸ Yujun Han,^{13,17} Hyeran Lee,¹⁹ Pinghua Li,¹⁴ Damon R. Lisch,²⁰ Sanzhen Liu,³ Zhijie Liu,⁶ Dawn Holligan Nagel,^{13,17} Maureen C. McCann,²¹ Phillip SanMiguel,²² Alan M. Myers,²³ Dan Nettleton,²⁴ John Nguyen,²⁵ Bryan W. Penning,^{15,21} Lalit Ponnala,²⁶ Kevin L. Schneider,²⁷ David C. Schwartz,²⁸ Anupma Sharma,²⁷ Carol Soderlund,²⁹ Nathan M. Springer,³⁰ Qi Sun,²⁶ Hao Wang,^{13,17} Michael Waterman,²⁵ Richard Westerman,²² Thomas K. Wolfgruber,²⁷ Lixing Yang,¹³ Yeisoo Yu,²⁹ Lifang Zhang,⁶ Shiguo Zhou,²⁸ Qihui Zhu,^{13,17} Jeffrey L. Bennetzen,¹³ R. Kelly Dawe,^{13,17} Jiming Jiang,¹⁹ Ning Jiang,³¹ Gernot G. Presting,²⁷ Susan R. Wessler,^{13,17} Srinivas Aluru,^{1,9,32} Robert A. Martienssen,⁶ Sandra W. Clifton,⁷ W. Richard McCombie,⁶ Rod A. Wing,⁸ Richard K. Wilson,^{7,33†}

We report an improved draft nucleotide sequence of the 2.3-gigabase genome of maize, an important crop plant and model for biological research. Over 32,000 genes were predicted, of which 99.8% were placed on reference chromosomes. Nearly 85% of the genome is composed of hundreds of families of transposable elements, dispersed nonuniformly across the genome. These were responsible for the capture and amplification of numerous gene fragments and affect the composition, sizes, and positions of centromeres. We also report on the correlation of methylation-poor regions with *Mu* transposon insertions and recombination, and copy number variants with insertions and/or deletions, as well as how uneven gene losses between duplicated regions were involved in returning an ancient allotetraploid to a genetically diploid state. These analyses inform and set the stage for further investigations to improve our understanding of the domestication and agricultural improvements of maize.

Maize (*Zea mays* ssp. *mays* L.) was domesticated over the past ~10,000 years from the grass teosinte in Central America (1) and has been subject to cultivation and selection ever since. Maize is an important model organism for fundamental research into the inheritance and functions of genes, the physical linkage of genes to chromosomes, the mechanistic relation between cytological crossovers and recombination, the origin of the nucleolus, the properties of telomeres, epigenetic silencing, imprinting, and transposition (2). Maize also is an important crop, yielding in the USA alone 12 billion (B = 10⁹) bushels of grain from ~86 million acres with a value of \$47 B [2008 data from (3)]. Over the last century, breeders have increased grain yields

eightfold (4), in part by harnessing heterosis (hybrid vigor), a universal, but poorly understood, phenomenon that can increase yields of hybrids by 15 to 60% relative to inbred parents (5).

The maize genome has undergone several rounds of genome duplication, including that of a paleopolyploid ancestor ~70 million years ago (mya) (6) and an additional whole-genome duplication event about 5 to 12 mya (7, 8), which distinguishes maize from its close relative, *Sorghum bicolor* (9). The 10 chromosomes of the maize genome are structurally diverse and have undergone dynamic changes in chromatin composition. The size of the maize genome has expanded dramatically (to 2.3 gigabases) over the last ~3 million years via a proliferation of

long terminal repeat retrotransposons (LTR retrotransposons) (10).

We sequenced the maize genome using a minimum tiling path of bacterial artificial chromosomes (BACs) ($n = 16,848$) and fosmid ($n = 63$) clones derived from an integrated physical and genetic map (11, 12), augmented by comparisons with an optical map (13). Clones were shotgun sequenced (four- to sixfold coverage), followed by automated and manual sequence improvement (14) of the unique regions only, which resulted in the B73 reference genome version 1 (B73 RefGen_v1).

We identified the full complement of maize transposable elements (TEs) accessible from B73 RefGen_v1, which includes active class II DNA TEs and an abundance of class I RNA TEs (15).

¹Center for Plant Genomics, Iowa State University, Ames, IA 50011, USA. ²Department of Agronomy, Iowa State University, Ames, IA 50011, USA. ³Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011, USA. ⁴Center for Carbon Capturing Crops, Iowa State University, Ames, IA 50011, USA. ⁵U.S. Department of Agriculture (USDA), North Atlantic Area, Robert Holley Center for Agriculture and Health, Ithaca, NY 14853, USA. ⁶Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA. ⁷The Genome Center at Washington University, St. Louis, MO 63108, USA. ⁸Arizona Genomics Institute, School of Plant Sciences and Department of Ecology and Evolutionary Biology, BIO5 Institute for Collaborative Research, University of Arizona, Tucson, AZ 85721, USA. ⁹Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011, USA. ¹⁰Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA. ¹¹School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA 99164, USA. ¹²Department of Botany, University of Florida, Gainesville, FL 32611, USA. ¹³Department of Genetics, University of Georgia, Athens, GA 30602, USA. ¹⁴Boyce Thompson Institute, Cornell University, Ithaca, NY 14853, USA. ¹⁵Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN 47907, USA. ¹⁶Université de Perpignan Via Domitia, CNRS, Perpignan, France. ¹⁷Department of Plant Biology, University of Georgia, Athens, GA 30602, USA. ¹⁸NimbleGen, Madison, WI 53711, USA. ¹⁹Department of Horticulture, University of Wisconsin–Madison, Madison, WI 53706, USA. ²⁰Department of Plant Biology, University of California, Berkeley, CA, 94720, USA. ²¹Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA. ²²Department of Horticulture and Landscape Architecture, Purdue University, West Lafayette, IN 47907, USA. ²³Department of Biochemistry, Biophysics, and Molecular Biology, Iowa State University, Ames, IA, 50011, USA. ²⁴Department of Statistics, Iowa State University, Ames, IA 50011, USA. ²⁵Departments of Mathematics, Biology, and Computer Science, University of Southern California, Los Angeles, CA 90089, USA. ²⁶Cornell University Computational Biology Service Unit, Cornell University, Ithaca, NY 14850, USA. ²⁷Molecular Biosciences and Bioengineering, University of Hawaii, Honolulu, HI 96822, USA. ²⁸Laboratory for Molecular and Computational Genomics, Department of Chemistry, Laboratory of Genetics, University of Wisconsin–Madison, Madison, WI 53706, USA. ²⁹BIO5 Institute for Collaborative Research, University of Arizona, Tucson, AZ 85721, USA. ³⁰Department of Plant Biology, University of Minnesota, St. Paul, MN 55108, USA. ³¹Department of Horticulture, Michigan State University, East Lansing, MI 48824, USA. ³²Indian Institute of Technology, Bombay, India. ³³Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA.

*These authors contributed equally to this work.

†These authors contributed equally to data production and analysis.

‡To whom correspondence should be addressed. E-mail: rwilson@wustl.edu

Almost 85% of the B73 RefGen_v1 consists of TEs (table S2). Indeed, the existence of TEs (16), as well as the first members of the *CACTA* (*Spm/En*), *hAT* (*Ac*), *PIF/Harbringer* and *Mutator* superfamilies, and MITE family (*Tourist*), were all initially discovered in maize (17). Further, both the existence and unparalleled abundance of LTR retrotransposons in plants were originally discovered in maize (18).

The B73 RefGen_v1 contains 855 families of DNA TEs that make up 8.6% of the genome; most of these (82%) were identified in this study (table S2) (14). The most complex of these superfamilies is *Mutator*, with dramatic variation in element sequence and size, including 262 Pack-MULEs (*Mutator*-like elements that contain gene

fragments) carrying fragments of 226 nuclear genes. About 40,000 nonredundant *Mu* insertion sites were amplified from *Mu*-active lines, sequenced, and mapped to B73 RefGen_v1. The nonuniformly distributed *Mu* insertion sites colocalize with gene-rich regions of the genome that have the highest rates of meiotic recombination per megabase (Fig. 1) (19). Like *Mu*, most maize DNA TEs (but not the *CACTA* elements) were enriched in the gene-rich, recombinationally active chromosome ends (Fig. 1 and fig. S1).

Helitrons, a class of DNA elements believed to transpose by a rolling-circle mechanism (20), are present in plants, animals, and fungi, but are particularly active, variable, and abundant in maize (21). Maize contains eight families of *Helitrons*

with a combined copy number of ~20,000, which are particularly active in gene fragment acquisition (table S2). In maize, we observed that *Helitrons* are located predominantly within gene-rich regions, whereas, in all previously studied plant and animal genomes, they are enriched in gene-poor regions (22, 23). LTR retrotransposons compose >75% of the B73 RefGen_v1 and are diverse. Most of the 406 families have fewer than 10 copies. LTR retrotransposons exhibited family-specific, nonuniform distributions along chromosomes, e.g., *Copia*-like elements are overrepresented in gene-rich euchromatic regions, whereas *Gypsy*-like elements are overrepresented in gene-poor heterochromatic regions (fig. S1) (24, 25). We observed more than 180 acquisitions of nuclear gene fragments inside LTR retrotransposons (table S2).

Protein-encoding and microRNA (miRNA) (26) genes were predicted from assembled or improved BAC contigs by a combination of evidence-based (27) and ab initio approaches, projected to B73 RefGen_v1, and subsequently filtered to a set of 32,540 protein-encoding and 150 miRNA genes (14) (fig. S2). Exon sizes of maize genes were similar to that of their orthologous genes in rice and sorghum, but maize genes contained more large introns because of insertion of repetitive elements (11, 28) (figs. S3 and S4 and tables S5 and S6). A comparative analysis with rice, sorghum, and *Arabidopsis* revealed similar numbers of gene families (14) (Fig. 2), of which a core set of 8494 families is shared among all four species, and of the 11,892 maize families, all but 465 are conserved with at least one other species. Species- and lineage-specific families point out potential inconsistencies between annotation projects, but also reflect genuine biological differences in gene inventories.

Because of the stringent criteria used for including genes in the filtered gene set (14), we expected to miss some genes. About 95% of a collection of 63,851 full-length maize cDNAs (fl-cDNAs) (29, 30) mapped to B73 RefGen_v1. On the basis of the ratio of fl-cDNA to supported genes in the filtered set, we estimated that this set accounts for at least 85% of all genes in the B73 RefGen_v1 (14).

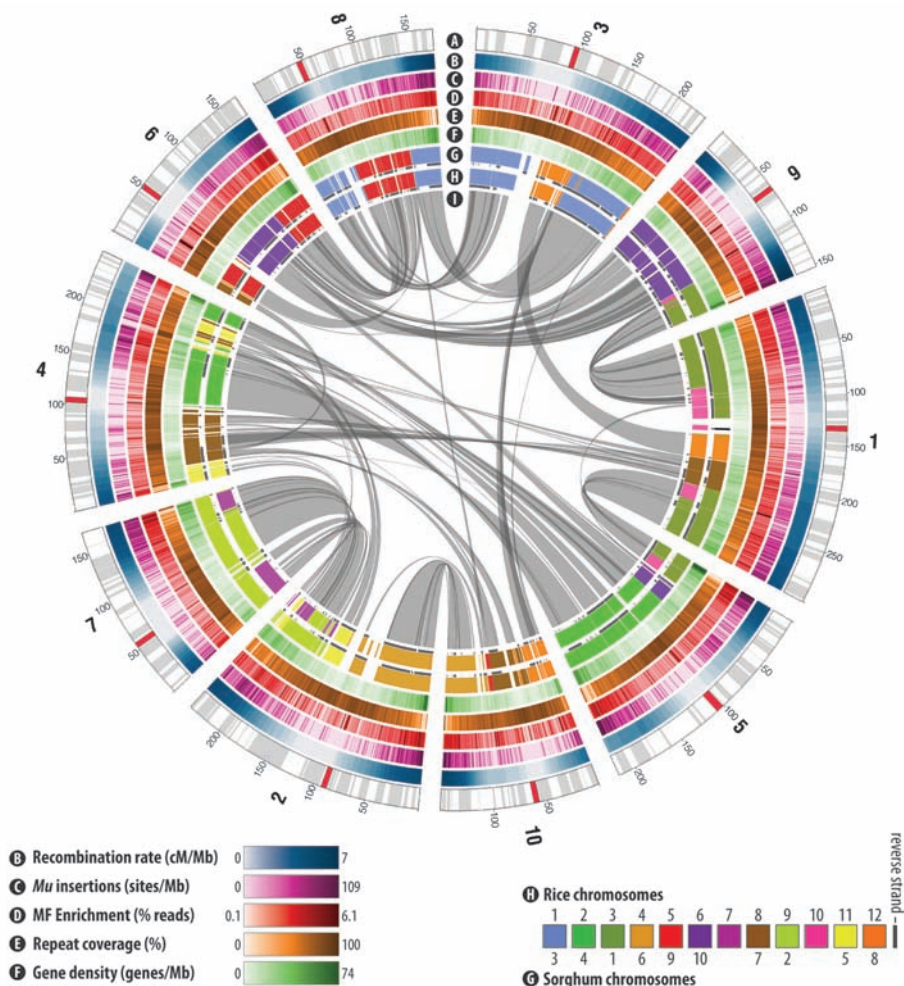


Fig. 1. The maize B73 reference genome (B73 RefGen_v1): Concentric circles show aspects of the genome. Chromosome structure (A). Reference chromosomes with physical fingerprint contigs (11) as alternating gray and white bands. Presumed centromeric positions are indicated by red bands (31); enlarged for emphasis. Genetic map (B). Genetic linkage across the genome, on the basis of 6363 genetically and physically mapped markers (14, 19). *Mu* insertions (C). Genome mappings of nonredundant *Mu* insertion sites (14, 19). Methyl-filtration reads (D). Enrichment and depletion of methyl filtration. For each nonoverlapping 1-Mb window, read counts were divided by the total number of mapped reads. Repeats (E). Sequence coverage of TEs with RepeatMasker with all identified intact elements in maize. Genes (F). Density of genes in the filtered gene set across the genome, from a gene count per 1-Mb sliding window at 200-kb intervals. Sorghum synteny (G) and rice synteny (H). Syntenic blocks between maize and related cereals on the basis of 27,550 gene orthologs. Underlined blocks indicate alignment in the reverse strand. Homoeology map (I). Oriented homoeologous sites of duplicated gene blocks within maize.

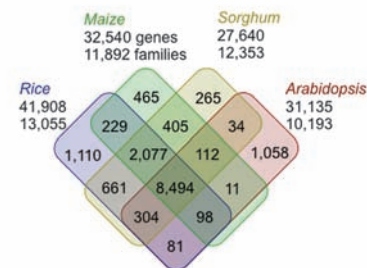


Fig. 2. Venn diagram showing unique and shared gene families between and among the three sequenced grasses (maize, rice, and sorghum) and the dicot, *Arabidopsis*.

The maximum rate of false-positive gene annotations was estimated by aligning ~112 million RNA-seq (transcriptome sequencing) reads from various tissues to the filtered gene set (14) (figs. S10 and S11). These experiments provided evidence for the transcription of ~91% of the genes in the filtered gene set (29,541 out of 32,540). Manual annotation of 200 randomly chosen genes from the filtered gene set indicated that only two are likely to be TE-derived. Additional manual annotation of smaller sets of selected genetically well-characterized genes (tables S8 to S10) indicated that the vast majority of genes and proteins predicted in the filtered gene set are mostly correct.

Maize centromeres were found to contain variable amounts of the tandem CentC satellite repeat and centromeric retrotransposon elements of maize (CRMs). On the basis of comparisons to B73 whole-genome shotgun data, we initially identified about half of the genome's CentC content (table S13). We captured additional CentC sequence by draft sequencing 101 centromeric repeat-containing BACs and anchoring them to the genetic and physical maps, thereby localizing all of the centromeres (31). We delineated the functional centromeres on the basis of their centromere-specific histone H3 (CENH3) (32) by using chromatin immunoprecipitation (ChIP) with an antibody against CENH3, followed by pyrosequencing. The centromere regions delineated in this way, although mostly incomplete, correlated with a high density of CentC and CRM1/CRM2/CRM3 repeats, but a number of these repeats also occurred outside of the functional centromeres (fig. S12). The CRM2 subfamily appears to be the centromeric repeat most closely associated with CENH3 in maize, as it is more enriched in the CENH3 chromatin fraction than CentC, CRM1, or CRM3 (table S13).

We traversed two centromeres (2 and 5) in their entirety and determined that they differ in size and CENH3 density (31). Because CRM elements have generated recombinants with distinct periods of activity (33, 34), we were able to demonstrate that the regional centromeres of maize are dynamic loci and that the CENH3 domain shifts over time (31).

To protect genome integrity, TEs are usually transcriptionally silenced (35) in part via the RNA-directed DNA methylation (RdDM) pathway, which requires an RNA-dependent RNA polymerase 2 (RDR2). When the maize homolog of RDR2 (36) is mutated, it alters the accumulation of transcripts from many characterized transposons, but unexpectedly, some TEs are down-regulated by loss of RDR2 function (37). In most plant genomes, genes are less densely methylated than heterochromatic TEs and other repeats. Consequently, ~2× coverage of the maize genome by methylation-filtered (MF) reads includes portions of ~95% of maize genes (38). Mapping MF reads (39) of maize and sorghum onto their respective genomes revealed species-specific distributions of heterochromatic DNA

methylation along the reference chromosomes (fig. S13, A and B). It is noteworthy that, in the sorghum genome, hypomethylated genes are largely excluded from the pericentromeric regions, whereas they are dispersed more widely in maize. Visual comparisons between sorghum and maize (14) revealed high levels of coalignment, including centromeres where centromeric repeats are undermethylated relative to the surrounding heterochromatin (39, 40) (fig. S13C). Thus, the B73 RefGen_v1 yields evidence that heavily methylated regions are more condensed during interphase.

Anchoring the B73 RefGen_v1 to a newly developed genetic map (19) revealed that rates of meiotic recombination per megabase are highest at the ends of the reference chromosomes and very low in the middle half of each chromosome surrounding the centromeres (Fig. 1) (19, 41). Although recombination occurs preferentially in genes (2) and gene density shows a similar distribution (Fig. 1), gene density does not fully explain the nonrandom distribution of recombination events, because a pronounced nonuniform distribution is still observed even when gene density is taken into consideration (19). Instead, epigenetic marks, including hypomethylation and histone modifications, are implicated in guiding both *Mu* insertion and meiotic recombination (19). Epigenetic processes have also been invoked to explain the observation that genomic imprinting contributes to the expression of thousands of genes in maize hybrids (42).

Maize exhibits extremely high levels of both phenotypic and genetic diversity. This genomic diversity was explored with both resequencing (41) and array-based comparative genomic hybridization between the B73 and Mo17 inbred lines (43). This revealed extensive structural variation, including hundreds of copy number variants (CNVs) and thousands of present-absent variants (PAVs). Many of the PAVs, including an ~2-Mb region on chromosome 6, contain intact, expressed single-copy genes that are present in one inbred genome but absent from the other. These haplotype-specific sequences may contribute to heterosis and the substantial degree of phenotypic variation among maize inbreds (43).

After a whole-genome duplication, the return to a genetically diploid state was associated with numerous chromosomal breakages and fusions, as shown by alignment to the genomes of sorghum and the more distantly related rice (Fig. 1 and fig. S14) (12). In contrast, sorghum has experienced relatively few interchromosomal rearrangements since its lineage split with rice (8); therefore, its chromosomal configuration closely resembles the ancestral state of maize's two subgenomes (12). Cosynteny of maize genes to common reference genes in rice or sorghum defined maize's duplicate regions (fig. S15). Although syntenic blocks cover 1832 Mb (~89% of the genome), individual gene losses were common and resulted in retention of only ~8110 genes as duplicate homoeologs (~25% of total

genes; ~30% having orthologs in rice and/or sorghum). On the basis of an analysis of GO (gene ontology) terms (14, 44) (table S15), retention of genes as duplicates is not random, e.g., retained duplicates are significantly enriched for transcription factors (>1.5-fold; *P* value = 7.6×10^{-22}) (table S15), as is also the case in rice (44) and *Arabidopsis* (45). An example of biased retention is the *CesA* family, in which all 10 ancestral sites were retained as duplicates (fig. S16) (46). Using the sorghum genome to project extant maize regions to ancestral chromosomes (14) revealed a strong bias for gene loss (fractionation) between sister regions (table S16 and fig. S17). Fractionation bias has been observed in other plant lineages and species (47–50).

Sites containing proximately duplicated paralogs tend to exist as single copies, or not at all, at corresponding homoeologous positions (table S18). Of the 1454 proximately duplicated paralogs identified (making up 3614 genes), only 126 (~9%) could be found at homoeologous positions (14). Of the remainder, 279 (19%) had a single paralog at the corresponding homoeologous site, and 1049 (72%) had no homoeologs.

Nearly identical paralogs (NIPs) are genes with pairwise alignments of ≥ 500 bp, $\geq 98\%$ identity, and $\geq 95\%$ coverage with other genes (51). Of maize-filtered genes, 2.5% (828 out of 32,540) were NIPs from 386 families, most of which have only two members (*n* = 349); the largest has nine members. Almost half (46%) of the NIP pairs had both members physically linked within 200 kb of each other, whereas in most of the remaining cases, the two members were distant from each other or on different chromosomes (fig. S18).

Just as cytogenetic and genetic maps (52) revolutionized research and crop improvement over the last century, the B73 maize reference sequence promises to advance basic research and to facilitate efforts to meet the world's growing needs for food, feed, energy, and industrial feed stocks in an era of global climate change. Findings derived from this genome sequence briefly summarized here are described in more detail in a series of companion papers (11, 13, 19, 22, 24–26, 30, 31, 37, 41–43, 46). Annotation data and browser are available at www.maizegenome.org.

References and Notes

- J. F. Doebley, B. S. Gaut, B. D. Smith, *Cell* **127**, 1309 (2006).
- J. L. Bennetzen, S. Hake, *Handbook of Maize: Genetics and Genomics* (Springer, New York, 2009).
- C. P. National Corn Growers Association, Table showing corn harvested, yield, production, mya price, and value, 1991–2008; <http://ncga.com/corn-production-trends>.
- A. F. Troyer, *Crop Sci.* **46**, 528 (2006).
- D. N. Duvick, *Science* **286**, 418 (1999).
- A. H. Paterson, J. E. Bowers, B. A. Chapman, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 9903 (2004).
- G. Blanc, K. H. Wolfe, *Plant Cell* **16**, 1667 (2004).
- Z. Swigonova et al., *Genome Res.* **14**, 1916 (2004).
- A. H. Paterson et al., *Nature* **457**, 551 (2009).
- P. SanMiguel, B. S. Gaut, A. Tikhonov, Y. Nakajima, J. L. Bennetzen, *Nat. Genet.* **20**, 43 (1998).

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. Seighe, 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

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