The B73 Maize Genome: Complexity, Diversity, and Dynamics

Patrick S. Schnable1,2,3,4, Doreen Ware5,6, Robert S. Fulton,7,†, Joshua C. Stein,4,†, Fusheng Wei8,†, Shiran Pasanen9, Chengzheng Liang,6, Jianwei Zhang,8, Lucinda Fulton,7, Tina A. Graves,7,†, Patrick Minx,5, Amy Deniese Reilly,9, Laura Courtney,5, Scott S. Kuchrowski,5, Chad Tomlinson,5, Cindy Strong,10, Kim Delehaunty,11, Carina Fronick,5, Bill Courtney,5, Susan M. Rock,7, Eddie Belter,6, Feiyu Du,12, Kyung Kim,13, Rachel M. Abbott,7, Andrew Kelly,13, Marilyn D. Peterson,7, Andrew M. Myers,23, Dan Nettleton,24, John Nguyen,25, Bryan W. Penning,15,21, Lalit Ponnala,26, Jeffrey A. Jeddeloh,18, Yujun Han,13,17, Hyeran Lee,19, Pinghua Li,14, Damon R. Lisch,20, Cristian Chaparro,16, Jer-Ming Chia,6, Jean-Marc Deragon,16, James C. Estill,13,17, Yan Fu,2,4, W. Brad Barbazuk,12, Regina S. Baucom,13, Thomas P. Brutnell,14, Nicholas C. Carpita,15, Shawn Leonard,7, Kevin Crouse,7, Kristi Collura,8, Dave Kudrna,8, Jennifer Currie,8, Ruifeng He,8, Angelina Angelova,8, Shamugam Rajasekar,14, Teri Mueller,14, Remi Lome,15, Gabriel Scara,1, Ara Ko,8, Krista Delaney,16, Marina Wissotzki,5, Georgina Lopez,5, David Campos,8, Michele Braidotti,8, Elizabeth Ashley,8, Wolfgang Golser,8, Hyeran Kim,8, Seunghee Lee,8, Jinke Lin,8, Zeljko Dujmic,8, W. Richard McCombie,6, Rod A. Wing,8, Richard K. Wilson7,33, Robert S. Fulton,7, Robert S. Fulton,7, Paul, MN 55108, USA.31Department of Horticulture, Michigan State University, East Lansing, MI 48824, USA.32Department of Electrical and Computer Engineering, University of Notre Dame, Notre Dame, IN 46556, USA.11School of Electrical Engineering and Computer Science, Washington University School of Medicine, St. Louis, MO 63110, USA.24Cornell University Computational Biology, University of Georgia, Athens, GA 30602, USA.25Departments of Mathematics, Biology, and Computer Science, Washington University School of Medicine, St. Louis, MO 63110, USA.12Department of Biochemistry, Biophysics, and Molecular Biology, Iowa State University, Ames, IA 50011, USA.13Department of Plant Biology, University of California, Berkeley, CA 94720, USA.14Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA.15Department of Horticulture and Landscape Architecture, Purdue University, West Lafayette, IN 47907, USA.16Laboratory of Genetics, University of Wisconsin-Madison, Madison, WI 53706, USA.17Department of Biochemistry, Biophysics, and Molecular Biology, Iowa State University, Ames, IA 50011, 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 Plant Sciences, University of Arizona, Tucson, AZ 85721, USA.22Department of Botany, University of Florida, Gainesville, FL 32611, USA.23Department of Genetics, University of Georgia, Athens, GA 30602, USA.24Boyle Thompson Institute, Cornell University, Ithaca, NY 14853, USA.25Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN 47907, USA.26Cornell University Computational Biology Service Unit, Cornell University, Ithaca, NY 14850, USA.27University of Perpignan Via Domitia, CNRS, Perpignan, France.28Department of Plant Biology, University of Georgia, Athens, GA 30602, USA.29BIO5 Institute for Collaborative Research, University of Arizona, Tucson, AZ 85721, USA.30Department of Statistics, Iowa State University, Ames, IA 50011, USA.21Departments of Mathematics, Biology, and Computer Science, University of Southern California, Los Angeles, CA 90089, USA.22Cornell University Computational Biology Service Unit, Cornell University, Ithaca, NY 14850, USA.23Molecular Biosciences and Bioengineering, University of Hawaii, Honolulu, HI 96822, USA.24Laboratory for Mole-}

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 transposable elements (TEs) and an abundance of class I RNA TEs and an abundance of class I RNA TEs, which can increase yields of hybrids (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).
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/Harbinger 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 co-localize 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 over-represented 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).
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 (CRM). 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 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 repeat sequences. 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 epigenetic marks, including hypomethylation and histone modifications, are implicated in guiding 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). Cosynthy of maize genes to common reference genes in rice or sorghum defined maize’s duplicate regions (fig. S15). Although syntetic 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; \( P = 7.6 \times 10^{-22} \)) (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, ~50% identity, and ≥50% 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.maizegdb.org.

References and Notes
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 genetic diversity (\(\delta\)) of maize (\(\delta = 0.39\)) 3.3 million single-nucleotide polymorphisms (SNPs) and indels (table S1) and found that, over- all, 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 \(\pi\) (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 base pairs (MbP) of low-copy sequence present in 13 or more lines in this study. Roughly 39% of the low-copy sequence was derived from introns and exons \((\delta)\), covering 32% of the total genomic fraction in the genome. We identified 3.3 million single-nucleotide polymorphisms (SNPs) and indels (table S1) and found that, overall, 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 \(\pi\) (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 base pairs (MbP) of low-copy sequence present in 13 or more lines in this study. Roughly 39% of the low-copy sequence was derived from introns and exons \((\delta)\), covering 32% of the total genomic fraction in the genome. We identified 3.3 million single-nucleotide polymorphisms (SNPs) and indels (table S1) and found that, overall, 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 \(\pi\) (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 base pairs (MbP) of low-copy sequence present in 13 or more lines in this study. Roughly 39% of the low-copy sequence was derived from introns and exons \((\delta)\), covering 32% of the total genomic fraction in the genome. We identified 3.3 million single-nucleotide polymorphisms (SNPs) and indels (table S1) and found that, overall, 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 \(\pi\) (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 base pairs (MbP) of low-copy sequence present in 13 or more lines in this study. Roughly 39% of the low-copy sequence was derived from introns and exons \((\delta)\), covering 32% of the total genomic fraction in the genome. We identified 3.3 million single-nucleotide polymorphisms (SNPs) and indels (table S1) and found that, overall, 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 \(\pi\) (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 base pairs (MbP) of low-copy sequence present in 13 or more lines in this study. Roughly 39% of the low-copy sequence was derived from introns and exons \((\delta)\), covering 32% of the total genomic fraction in the genome. We identified 3.3 million single-nucleotide polymorphisms (SNPs) and indels (table S1) and found that, overall, 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 \(\pi\) (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