Title
Improved maize reference genome with single-molecule technologies

Permalink
https://escholarship.org/uc/item/7c45d4qb

Journal
Nature, 546(7659)

ISSN
0028-0836

Authors
Jiao, Yinping
Peluso, Paul
Shi, Jinghua
et al.

Publication Date
2017-06-01

DOI
10.1038/nature22971

Peer reviewed
Improved maize reference genome with single-molecule technologies

Yinping Jiao1, Paul Peluso2, Jinghua Shi3, Tiffany Liang3, Michelle C. Stitzer4, Bo Wang1, Michael S. Campbell1, Joshua C. Stein1, Xuehong Wei1, Chen–Shan Chin2, Katherine Guill3, Michael Regulski1, Sunita Kumar1, Andrew Olson1, Jonathan Gent6, Kevin L. Schneider7, Thomas K. Wolfrugber7, Michael R. May8, Nathan M. Springer3, Eric Antoniou1, W. Richard McCombie1, Gernot G. Presting7, Michael McMullen5, Jeffrey Ross–Ibarra10, R. Kelly Dawe6, Alex Haste1, David R. Rank2 & Doreen Ware1,11

Complete and accurate reference genomes and annotations provide fundamental tools for characterization of genetic and functional variation1. These resources facilitate the determination of biological processes and support translation of research findings into improved and sustainable agricultural technologies. Many reference genomes for crop plants have been generated over the past decade, but these genomes are often fragmented and missing complex repeat regions2. Here we report the assembly and annotation of a reference genome of maize, a genetic and agricultural model species, using single-molecule real-time sequencing and high-resolution optical mapping. Relative to the previous reference genome3, our assembly features a 52-fold increase in contig length and notable improvements in the assembly of intergenic spaces and centromeres. Characterization of the repetitive portion of the genome revealed more than 130,000 intact transposable elements, allowing us to identify transposable element lineage expansions that are unique to maize. Gene annotations were updated using 111,000 full-length transcripts obtained by single-molecule real-time sequencing4. In addition, comparative optical mapping of two other inbred maize lines revealed a prevalence of deletions in regions of low gene density and maize lineage-specific genes.

Maize is the most productive and widely grown crop in the world, as well as a foundational model for genetics and genomics5. An accurate genome assembly for maize is crucial for all forms of basic and applied research, which will enable increases in yield to feed the growing world population. The current assembly of the maize genome, based on Sanger sequencing, was first published in 2009 (ref. 3). Although this initial reference enabled rapid progress in maize genomics, the original assembly is composed of more than 100,000 small contigs, many of which are arbitrarily ordered and oriented, markedly complicating detailed analysis of individual loci6 and impeding investigation of intergenic regions crucial to our understanding of phenotypic variation7,8 and genome evolution9,10.

Here we report a vastly improved de novo assembly and annotation of the maize reference genome (Fig. 1). On the basis of 65 × single-molecule real-time sequencing (SMRT) (Extended Data Fig. 1), we assembled the genome of the maize inbred line B73 into 2,958 contigs, in which half of the total assembly is made up of contigs larger than 1.2 Mb (Table 1, Extended Data Fig. 2a). The assembly of the long reads was then integrated with a high-quality optical map (Extended Data Fig. 1, Extended Data Table 1) to create a hybrid assembly consisting of 625 scaffolds (Table 1). To build chromosome-level super-scaffolds, we combined the hybrid assembly with a minimum tiling path generated from the bacterial artificial chromosomes (BACs)11 and a high-density genetic map12 (Extended Data Fig. 2b). After gap-filling and error correction using short sequence reads, the total size of maize B73 RefGen_v4 pseudomolecules was 2,106 Mb. The new reference assembly has 2,522 gaps, of which almost half (n = 1,115) have optical

Figure 1 | Genome assembly layout. a, Workflow for genome construction. b, Ideograms of maize B73 version 4 reference pseudomolecules. The top track shows positions of 2,522 gaps in the pseudomolecules, including 1,115 gaps in which the lengths were estimated using optical genome maps (orange), whereas the remainder (purple) have undetermined lengths. More than half of the assembly is constituted of contigs longer than 1 Mb, which are shown as light grey bars in the bottom track.

1Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA. 2Pacific Biosciences, Menlo Park, California 94025, USA. 3BioNano Genomics, San Diego, California 92121, USA. 4Department of Plant Sciences and Center for Population Biology, University of California, Davis, California 95616, USA. 5USDA-ARS, Plant Genetics Research Unit, Columbia, Missouri 65211, USA. 6University of Georgia, Athens, Georgia 30602, USA. 7Department of Molecular Biosciences and Bioengineering, University of Hawaii, Honolulu, Hawaii 96822, USA. 8Department of Evolution and Ecology, University of California, Davis, California 95616, USA. 9Department of Plant Biology, University of Minnesota, St Paul, Minnesota 55108, USA. 10Department of Plant Sciences, Center for Population Biology, and Genome Center, University of California, Davis, California 95616, USA. 11USDA-ARS, NEA Robert W. Holley Center for Agriculture and Health, Cornell University, Ithaca, New York 14853, USA.
Our assembly made substantial improvements in the gene space including resolution of gaps and misassemblies and correction of order and orientation of genes. We also updated the annotation of our new assembly, resulting in consolidation of gene models (Extended Data Fig. 4). Newly published full-length cDNA data improved the annotation of alternative splicing by more than doubling the number of alternative transcripts from 1.6 to 3.3 per gene (Extended Data Fig. 5a), with about 70% of genes supported by the full-length transcripts. Our reference assembly also vastly improved the coverage of regulatory sequences, decreasing the number of genes exhibiting gaps in the 3-kb region(s) flanking coding sequence from 20% to <1% (Extended Data Fig. 5b). The more complete sequence enabled notable improvements in the annotation of core promoter elements, especially the TATA-box, CCAAT-box, and Y patch motifs (Supplementary Information). Quantitative genetic analyses have shown that polymorphisms in regulatory regions explain a substantial majority of the genetic variation for many phenotypes, suggesting that the new reference will markedly improve our ability to identify and predict functional genetic variation.

After its divergence from Sorghum, the maize lineage underwent genome doubling followed by diploidization and gene loss. Previous work showed that gene loss is biased towards one of the parental genomes, but our new assembly and annotation instead suggest that 56% of syntenic sorghum orthologues map uniquely to the dominant maize subgenome (designated A, total size 1.16 Gb), whereas only 24% map uniquely to subgenome B (total size 0.63 Gb). Gene loss in maize has primarily been considered in the context of polyploidy and functional redundancy, but we found that despite its polyploidy, maize has lost a larger proportion (14%) of the 22,048 ancestral gene orthologues than any of the other four grass species evaluated to date (Sorghum, rice, Brachypodium distachyon and Setaria italica; Extended Data Fig. 6). Nearly one-third of these losses are specific to maize, and analysis of a restricted high-confidence set revealed enrichment for genes involved in biotic and abiotic stresses (Extended Data Table 2), for example, NB-ARC domain disease-resistance genes and the serpin protease inhibitor involved in pathogen defence and programmed cell death.

Transposable elements were first reported in maize and have since been shown to have important roles in shaping genome evolution and gene regulatory networks of many species. Most of the maize
genome is derived from transposable elements\textsuperscript{3,21}, and careful study of a few regions has revealed a characteristic structure of sequentially nested retrotransposons\textsuperscript{1,22} and the effect of deletions and recombination on retrotransposon evolution\textsuperscript{23}. In the annotation of the original maize assembly, however, fewer than 1% of long terminal repeat (LTR) retrotransposon copies were intact\textsuperscript{24}. By applying a new homology-independent annotation pipeline to our assembly (Extended Data Table 3), we identified 1,268 Mb (130,604 copies) of structurally intact retrotransposons, of which 661 Mb (70,035 copies) are nested retrotransposon copies disrupted by the insertion of other transposable elements, 8.7 Mb (14,041 copies) are DNA terminal inverted repeat transposons, and 76 Mb (21,095 copies) are helitrons. To understand the evolutionary history of maize LTR retrotransposons, we also applied our annotation pipeline to the sorghum reference genome, and used reverse transcriptase protein domain sequences that were accessible owing to the improved assembly of the internal protein coding domains of maize LTR retrotransposons to reconstruct the phylogeny of maize and sorghum LTR retrotransposon families. Despite a higher overall rate of diversification of LTR transposable elements in the maize lineage consistent with its larger genome size, differences in LTR retrotransposon content between genomes were primarily the result of marked expansion of distinct families in both lineages (Fig. 2).

Maize exhibits tremendous genetic diversity\textsuperscript{25}, and both nucleotide polymorphisms and structural variations have important roles in its phenotypic variation\textsuperscript{12,26}. However, genome-wide patterns of structural variation in plant genomes are difficult to assess\textsuperscript{27}, and previous efforts have relied on short-read mapping, which misses the vast majority of variation in plant genomes are difficult to assess\textsuperscript{27}, and previous efforts have relied on short-read mapping, which misses the vast majority of variation in plant genomes. More than 90% of the indels were unique to one inbred or the other, indicating a high level of structural diversity in maize. As short-read sequence data are available from both Ki11 and W22 (ref. 10), we analysed 1,451 of the largest (>10 kb) deletions and found that 1,083 were supported by a clear reduction in read depth (Fig. 3c). The confirmed deletions occurred in regions of low gene density (4.4 genes per megabase compared to a genome-wide average of 18.7 genes per megabase). One-third (83 out of 257) of the genes missing in Ki11 or W22 lack putative orthologues in all four grasses (rice, sorghum, Brachypodium and Setaria), consistent with previous data\textsuperscript{28}.

Although maize is often considered to be a large-genome crop, most major food crops have even larger genomes with more complex repeat landscapes\textsuperscript{2}. Our improved assembly of the B73 genome, generated using single-molecule technologies, demonstrates that additional assemblies of other maize inbred lines and similar high-quality assemblies of other repeat-rich and large-genome plants are feasible. Further high-quality assemblies will in turn extend our understanding of the genetic diversity that forms the basis of the phenotypic diversity in maize and other economically important plants.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 27 September 2016; accepted 14 May 2017.**

**Published online 12 June 2017.**

---

1. Edwards, D., Batley, J. & Snowden, R. J. Accessing complex crop genomes with next-generation sequencing. *Theor. Appl. Genet.* **126**, 1–11 (2013).
2. Morris, P. L., Buckler, E. S. & Ross-Ibarra, J. Crop genomics: advances and applications. *Nat. Rev. Genet.* **13**, 85–96 (2011).
3. Schnable, P. S. et al. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115 (2009).
4. Wang, B. et al. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* **7**, 11708 (2016).
5. Hake, S. & Ross-Ibarra, J. Genetic, evolutionary and plant breeding insights from the domestication of maize. *eLife* **4**, (2015).
6. Fouquet, R. et al. Maize rough endosperm3 encodes an RNA splicing factor required for endosperm cell differentiation and has a nonautonomous effect on embryo development. *Plant Cell* **23**, 4280–4297 (2011).
METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Whole-genome sequencing using SMRT technology. DNA samples for SMRT sequencing were prepared using maize imbed line B73 from NCPRIS (PI550473), grown at University of Missouri. Seeds of this line were deposited at NCPRIS (tracking number P167/128). Etiolated seedlings were grown for 4–6 days in Pro-Mix BX (Scotts Miracle-Gro) to minimize chloroplast DNA. Batches of ~10 g were snap-frozen in liquid nitrogen. DNA was extracted following the PacBio protocol ‘Preparing Arabidopsis Genomic DNA for Size-Selected ~20 kb SMRTbell Libraries’ (http://www.pacb.com/wp-content/uploads/2015/09/Shared-Protocol-Preparing-Arabidopsis-DNA-for-20-kb-SMRTbell-Libraries.pdf).

Genomic DNA was sheared to a size range of 15–40 kb using either G-tubes (Covaris) or a Megaruptor device (Diagenode), and enzymatically repaired and converted into SMRTbell template libraries as recommended by Pacific Biosciences. In brief, hairpin adapters were ligated, after which the remaining damaged DNA fragments and those without adapters at both ends were eliminated by digestion with exonucleases. The resulting SMRTbell templates were size-selected by Blue Pippin electrophoresis (Sage Sciences) and templates ranging from 15 to 50 kb were sequenced on a PacBio RS II instrument using P6-C4 sequencing chemistry. To acquire long reads, all data were collected as either 5- or 6-h sequencing videos.

Construction of optical genome maps using the Irys system. High-molecular mass genomic DNA was isolated from 3 g of young ear tissue after fixing with 2% formaldehyde. Nuclei were purified and lysed in embedded agarose as previously described41. DNA was labelled at Nt.BspQI sites using the IrysPrep kit. Molecules collected from BioNano chips were de novo assembled as previously described31 using ‘optArgument_human’.

De novo assembly of the genome sequencing data. De novo assembly of the long reads from SMRT Sequencing was performed using two assemblers: the Celera Assembler PBrC–MHAP pipeline31 and Falcon34 with different parameter settings. Quiver from SMRT Analysis v2.3.0 was used to polish base calling of contigs. The three independent assemblies were evaluated by aligning with the optical genome maps.

Contamination of contigs by bacterial and plasmid genomes was eliminated using the NCBI GenBank submission system35. Curation of the assembly, including resolution of conflicts between the contigs and the optical map and removal of redundancy at the edges of contigs, is described in the Supplementary Information.

Hybrid scaffold construction. To create hybrid scaffolds, curated sequence contigs and optical maps were aligned and merged with ReAligner32 (P < 10⁻¹¹). These initial hybrid scaffolds were aligned again to the sequence contigs using a less stringent P value (1 × 10⁻⁸), and those contigs not previously merged were added if they aligned over 50% of their length and without overlapping previously merged contigs, thereby generating final hybrid scaffolds.

Pseudomolecule construction. Sequences from BACs on the physical map that were used to build the maize v3 pseudomolecules were aligned to contigs using MUMMER package36 with the following parameter setting: ‘-l(minimum length of a single match) 100 -c(minimum length of a cluster of matches) 1000’.

To only use unique hits as markers, alignment hits were filtered with the following parameters: ‘-l(minimum alignment identity) 98 -l(minimum alignment length) 10000’. Scaffolds were then ordered and oriented into pseudochromosomes using the order of BACs as a guide. For quality control, we mapped the SNP markers from a genetic map built via single-nucleotide polymorphism (SNP) markers from a genetic map built from an intermated maize recombinant inbred line population (Mo17 × B73)32. Contigs with markers not located in pseudochromosomes from the optical map were placed into the AGP (A Golden Path) using the genetic map.

Further polishing of pseudomolecules. Raw pseudomolecules were subjected to gap filling using Phellybe (maxTrim = 0, -minReads = 2) and polished again using Quiver (SMRT Analysis v2.3.0). To increase the accuracy of the base calls, we performed two lanes of sequencing on the same genomic DNA sample (library size = 450 bp) using Illumina 2500 Rapid run, which generated about 100-fold 2 × 250 paired-end (PE) data. Reads were aligned to the assembly using BWA-mem37. Sequence error correction was performed with the Pilon pipeline48, after aligning reads with BWA-mem37 and parsing with SAMtools49, using sequence and alignment quality scores above 20.

Annotation. For comprehensive annotation of transposable elements, we designed a structural identification pipeline incorporating several tools, including LTRharvest40, LTRigdet41, SINE-Finder 42, MGEScan-non-LTR43, MITE-hunter44, HelitronScanner45, and others (details in Supplementary Information). The scripts, parameters, and intermediate files of each transposable element superfamily are available at https://github.com/mcssttzer/maze_v4_TE_annotation.

The MAKER-P pipeline was used to annotate protein-coding genes46, integrating ab initio prediction with publicly available evidence from full-length cDNA47, de novo assembled transcripts from short-read mRNA sequencing (mRNA-seq)48, isoform-sequencing (Iso-Seq) full-length transcripts44, and proteins from other species. The gene models were filtered to remove transposons and low-confidence predictions. Additional alternative transcript isoforms were obtained from the Iso-Seq data. Further details on annotations, core promoter analysis, and comparative phylogenomics are described in Supplementary Information.

STUDY DEPOSITION

Sequencing data have been deposited at the NCBI under BioProject number PRJNA10769 and BioSample number SAMN04296295. PacBio whole-genome sequencing data and Illumina data were deposited in the NCBI SRA database under accessions SRX1472849 and SRX1452310, respectively. The GenBank accession number of the genome assembly and annotation is LPUQ0000000. A genome browser including genome feature tracks and ftp is available from Gramene: http://ensembl.gramene.org/Zea_mays/Info/Index. All other data are available from the corresponding author upon reasonable request.

1. Steinbiss, S., Willhoeft, U., Gremme, G. & Kurtz, S. Fine-grained annotation and de novo assembly of the maize genome using Bowtie2 (ref. 51). Bioinformatics 31, 2508–2517 (2015).
2. 19. Nakashima, S. et al. Versatile and open software for comparing large genomes. Genome Biol. 5, R12 (2004).
3. 20. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1756–1764 (2009).
4. 21. Watanabe, B. J. & Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9, e112963 (2014).
5. 22. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
6. 23. Ellingshaus, D., Kurtz, S. & Wiitholz, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9, 19 (2008).
7. 24. Steinbiss, S., Willhoeft, U., Gremme, G. & Kurtz, S. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Res. 37, 7002–7013 (2009).
8. 25. Venken, T. et al. Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. Plant Cell 23, 3117–3128 (2011).
9. 26. Rho, M. & Tang, H. MGEscan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes. Nucleic Acids Res. 37, e142 (2009).
10. 27. Harms, P. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res. 38, e199 (2010).
45. Xiong, W., He, L., Lai, J., Dooner, H. K. & Du, C. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. Proc. Natl Acad. Sci. USA 111, 10263–10268 (2014).
46. Campbell, M. S. et al. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. 164, 513–524 (2014).
47. Soderlund, C. et al. Sequencing, mapping, and analysis of 27,455 maize full-length cDNAs. PLoS Genet. 5, e1000740 (2009).
48. Law, M. et al. Automated update, revision, and quality control of the maize genome annotations using MAKER-P improves the B73 RefGen_v3 gene models and identifies new genes. Plant Physiol. 167, 25–39 (2015).
49. Mostovoy, Y. et al. A hybrid approach for de novo human genome sequence assembly and phasing. Nat. Methods 13, 587–590 (2016).
50. Cao, H. et al. Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. Gigascience 3, 34 (2014).
51. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).
Extended Data Figure 1 | Summary of data generated for genome construction. a, Size distribution of single molecules for the optical maps. A total of 150 Gb (~60-fold coverage) of single-molecule raw data from BioNano chips was collected for map construction. The N50 of the single molecules was ~261 kb, and the label density was 11.6 per 100 kb. After assembly, the total size of the map reached 2.12 Gb with an N50 of 2.47 Mb. b, Length distribution of SMRT sequencing reads. Sequencing of 212 P6-C4 SMRT cells on the PacBio platform generated ~65-fold depth-of-coverage of the nuclear genome. Read lengths averaged 11.7 kb, with reads above 10 kb providing 53-fold depth-of-coverage. c, The accuracy of SMRT sequencing from a representative run. The sequencing error rate was estimated at 10% from the alignment with the maize B73 RefGen_v3 by BLASR. d, Plot of the fraction of alignable data per run (alignable bases/total bases per chip) versus total raw bases (per chip) for each B73 sequencing run. As the plot shows, the trend in the data suggests that as the overall per run raw base yield increases, the fraction of alignable bases decreases. This is owing to the fact that in all runs, a subset of the zero-mode waveguide (ZMWs) will initially have more than one active sequencing enzyme in the observation field at the start of the sequencing run. A ZMW with more than one active polymerase will create unalignable bases while the two polymerases are simultaneously synthesizing DNA and yield a ‘merged sequencing signal from two independent polymerases’. As the loading of a chips increases (yield of bases), the probability of having two or more polymerases in a single ZMW increases.
Extended Data Figure 2 | Construction of pseudomolecules. a, Summary of the three assembly sets. b, How the scaffolds were ordered according to the order of the BACs. c, Size distribution of gaps in the pseudomolecules estimated using the optical map.
Extended Data Figure 3 | Quality assessment and comparison of the assembly in centromere and telomere regions in maize B73 RefGen_v3 and v4. 

a, Quality assessment of centromere and telomere using optical genome map. b, Locations of centromeres on pseudomolecules defined by ChIP–seq in the B73 RefGen_v3 and v4. c, Telomere repeats found in the B73 RefGen_v4 pseudomolecules.
Extended Data Figure 4 | Details of the gene annotation of maize B73 RefGen_v4. a, The pipeline used to characterize high confidence gene models. b, Summary of B73 RefGen_v4 protein-coding gene annotation, and comparison with RefGen_v3 annotation.
Extended Data Figure 5 | Improvement of the annotation of alternative splicing and completeness of regulatory regions of maize RefGen_v4 genes. 

a. Number of transcripts of each gene in v3 and v4 annotation. b. Percentages of genes with gaps in flanking regions in the v3 and v4 annotations.
Extended Data Figure 6 | Comparative analysis of the maize B73 RefGen_v4 genes with other grasses. a. Species-membership in orthologue sets, giving counts and percentage of orthologue sets of which each species is a member. Numbers in parentheses give the percentage of orthologue sets with membership of all species and versions within the clade. na, not applicable. b. Venn diagram showing overlap of 6,539 orthologue sets rooted in the Poaceae (true grasses) that are deficient in gene membership among five species.
Extended Data Figure 7 | Structural variation characterized from the Ki11 and W22 optical maps.
Extended Data Table 1 | Summary of the optical maps of three maize lines

| Length Bin  | # of Maps   | Quantity (Mb) | Bin proportion (% by mass) |
|-------------|-------------|---------------|--------------------------|
| 10–500 kb   | 311         | 102,462       | 5%                       |
|             | 675         | 213,642       | 10%                      |
|             | 540         | 179,105       | 7%                       |
| 500–1000 kb | 323         | 237,117       | 11%                      |
|             | 644         | 465.86        | 21%                      |
|             | 710         | 526.351       | 21%                      |
| 1000–2000 kb| 341         | 486.497       | 23%                      |
|             | 573         | 805.219       | 36%                      |
|             | 606         | 850.356       | 34%                      |
| >2000 kb    | 378         | 1293.607      | 61%                      |
|             | 256         | 731.371       | 33%                      |
|             | 331         | 974.819       | 39%                      |
### Extended Data Table 2 | Overrepresented protein domains in sorghum genes that lack orthologues in maize but are conserved in syntenic positions in other grasses

| Missing orthologs (n=668)† | Background (n=21,881)* | Pfam description | pval       | qval       |
|---------------------------|------------------------|-----------------|------------|------------|
| PF00646                  | 24                     | 162             | F-box domain     | 1.57E-10 | 5.86E-08  |
| PF03478                  | 11                     | 37              | DUF295       | 8.21E-09 | 1.53E-06  |
| PF07893                  | 6                      | 8               | DUF1668      | 2.10E-08 | 2.62E-06  |
| PF00931                  | 19                     | 146             | NB-ARC domain | 1.08E-07 | 1.01E-05  |
| PF07762                  | 7                      | 16              | DUF1618      | 2.16E-07 | 1.61E-05  |
| PF00079                  | 4                      | 10              | Serpin (serine protease inhibitor) | 1.56E-04 | 9.73E-03  |
| PF01754                  | 4                      | 11              | A20-like zinc finger | 2.39E-04 | 1.26E-02  |
| PF11443                  | 3                      | 5               | DUF2828      | 2.71E-04 | 1.26E-02  |
| PF01428                  | 4                      | 14              | AN1-like Zinc finger | 6.75E-04 | 2.80E-02  |
| PF12274                  | 3                      | 7               | DUF3615      | 9.04E-04 | 3.16E-02  |
| PF10266                  | 2                      | 2               | Hereditary spastic paraplegia protein strumpellin | 9.31E-04 | 3.16E-02  |
| PF08370                  | 4                      | 16              | Plant PDR ABC transporter associated | 1.17E-03 | 3.64E-02  |

*High-confidence sorghum genes with syntenic orthologues in rice, Brachypodium or Setaria outgroup species.
†Subset of background with no annotated orthologues in either maize v3 or v4 reference assemblies, have <50% LASTZ alignment coverage with v4, and fall within synteny blocks that map to singular assembly contigs in both the A and B subgenomes of maize. Only significantly enriched cases are shown, based on hypergeometric distribution followed by FDR correction.
Extended Data Table 3 | Structural annotation of transposable elements

| Order | Superfamily | Copies | Total size (bp) | Percentage of the genome assembly |
|-------|-------------|--------|----------------|-----------------------------------|
| LTR   |             | 136,604| 1,267,951,839  | 59.98%                            |
|       | RLC         | 45,032 | 386,862,053    | 18.30%                            |
|       | RLG         | 73,021 | 737,341,028    | 34.88%                            |
|       | RLX         | 18,551 | 143,748,758    | 6.80%                             |
| SINE  |             | 915    | 293,390        | 0.01%                             |
|       | RST         | 915    | 293,390        | 0.01%                             |
| LINE  |             | 65     | 121,583        | 0.01%                             |
|       | RIL         | 36     | 84,796         | 0.00%                             |
|       | RIT         | 29     | 36,787         | 0.00%                             |
| Helitron |       | 21,095 | 76,039,832    | 3.60%                             |
|       | DHH         | 21,095 | 76,039,832    | 3.60%                             |
| TIR   |             | 14,041 | 8,712,629      | 0.41%                             |
|       | DTA         | 5,646  | 3,265,936      | 0.15%                             |
|       | DTC         | 1,178  | 1,874,329      | 0.09%                             |
|       | DTH         | 5,136  | 1,418,803      | 0.07%                             |
|       | DTM         | 1,246  | 1,988,819      | 0.09%                             |
|       | DTT         | 835    | 164,742        | 0.01%                             |
| TOTAL |             | 184,067| 1,352,997,690  | 64.00%                            |