OPEN
doi:10.1038/nature22971

Improved maize reference genome with single-molecule technologies

Yinping Jiao[1], Paul Peluso[2], Jinghua Shi[3], Tiffany Liang[3], Michelle C. Stitzer[4], Bo Wang[1], Michael S. Campbell[1], Joshua C. Stein[1], Xuehong Wei[1], Chen-Shan Chin[2], Katherine Guill[5], Michael Regulski[1], Sunita Kumari[1], Andrew Olson[1], Jonathan Gent[6], Kevin L. Schneider[7], Thomas K. Wolfgruber[7], Michael R. May[8], Nathan M. Springer[9], Eric Antoniou[1], W. Richard McCombie[1], Gernot G. Presting[7], Michael McMullen[5], Jeffrey Ross-Ibarra[10], R. Kelly Dawe[6], Alex Hastie[3], David R. Rank[2] & Doreen Ware[1,11]

Complete and accurate reference genomes and annotations provide fundamental tools for characterization of genetic and functional variation[1]. 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 regions[2]. 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 genome[3], 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 sequencing[4]. 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 genomics[5]. 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[1], 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 loci[6] and impeding investigation of intergenic regions crucial to our understanding of phenotypic variation[7,8] and genome evolution[9,10].

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 map[12] (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

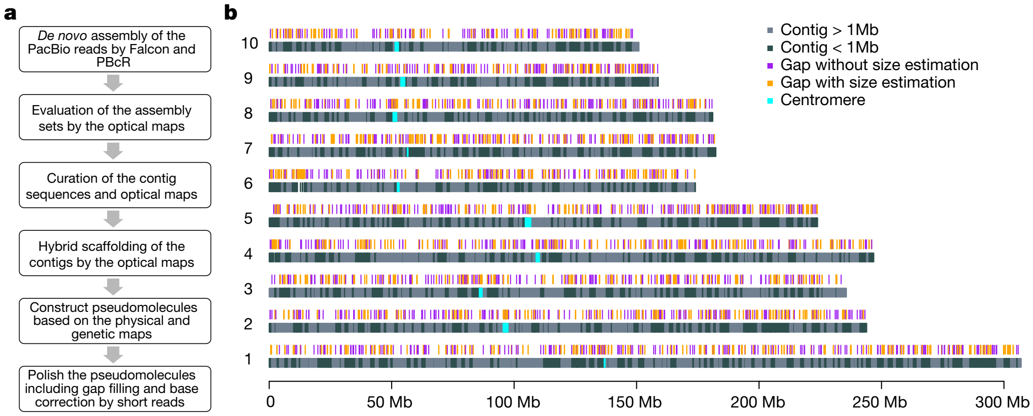


Figure 1 | Genome assembly layout. a, Workflow for genome construction. **b**, Ideograms of maize B73 version 4 reference pseudomolecules. The top track shows positions of 2,522 gaps in the pseudomolecules, including 1,115 gaps in which the lengths were estimated using optical genome maps (orange), whereas 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.

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA. [2]Pacific Biosciences, Menlo Park, California 94025, USA. [3]BioNano Genomics, San Diego, California 92121, USA. [4]Department of Plant Sciences and Center for Population Biology, University of California, Davis, Davis, California 95616, USA. [5]USDA-ARS, Plant Genetics Research Unit, Columbia, Missouri 65211, USA. [6]University of Georgia, Athens, Georgia 30602, USA. [7]Department of Molecular Biosciences and Bioengineering, University of Hawaii, Honolulu, Hawaii 96822, USA. [8]Department of Evolution and Ecology, University of California, Davis, California 95616, USA. [9]Department of Plant Biology, University of Minnesota, St Paul, Minnesota 55108, USA. [10]Department of Plant Sciences, Center for Population Biology, and Genome Center, University of California, Davis, California 95616, USA. [11]USDA-ARS, NEA Robert W. Holley Center for Agriculture and Health, Cornell University, Ithaca, New York 14853, USA.

Table 1 | Assembly statistics of the maize B73 RefGen_v4 genome

	Number of contigs (scaffolds)	Mean length (Mb)	N50 size (Mb)	Maximum length (Mb)	Total assembly length (Mb)
Original optical maps	1,342	1.57	2.47	12.43	2,107
Original contigs from sequence assembly	3,303	0.64	1.04	5.65	2,105
Curated optical maps	1,356	1.56	2.47	12.47	2,114
Curated contigs from sequence assembly	2,958	0.71	1.18	7.26	2,104
Optical maps in hybrid scaffolds	1,287	1.62	2.49	12.47	2,080
Contigs in hybrid scaffolds	2,696	0.77	1.19	7.26	2,075
Hybrid scaffolds	356	5.97	9.73	38.53	2,075
Hybrid scaffolds and non-scaffolded contigs	625	3.45	9.56	38.53	2,105

Table 2 | Structural variations from optical maps of two maize lines

	Ki11 map versus B73 RefGen_v4	W22 maps versus B73 RefGen_v4
Total size of genome map (Mb)	2,216	2,280
Map aligned to reference genome (Mb)	722	893
Reference genome covered by map (Mb)	694	861
Region in B73 with insertion and deletion (Mb)	223	221
Ratio of region with insertion and deletion (%)	32.15	25.67
Number of insertions	1,794	1,614
Average insertion size (bp)	21,510	21,470
Number of deletions	1,701	1,597
Average deletion size (bp)	18,340	20,120
Number of deletion regions potentially affecting genes	636	621

map coverage, giving an estimated mean gap length of 27 kb (Extended Data Fig. 2c). The new maize B73 reference genome has 240-fold higher contiguity than the recently published short-read genome assembly of maize cultivar PH207 (contig N50: 1,180 kb versus 5 kb)[13].

Comparison of the new assembly to the previous BAC-based maize reference genome assembly revealed more than 99.9% sequence identity and a 52-fold increase in the mean contig length, with 84% of the BACs spanned by a single contig from the long reads assembly. Alignment of chromatin-immunoprecipitation followed by sequencing (ChIP–seq) data for the centromere-specific histone H3 (CENH3)[14] revealed that centromeres are accurately placed and largely intact. Several previously identified[15] megabase-sized mis-oriented pericentromeric regions were also corrected (Extended Data Fig. 3a, b). Moreover, the ends of the chromosomes are properly identified on 14 out of the 20 chromosome arms based on the presence of tandem telomeric repeats and knob 180 sequences (Extended Data Fig. 3a, c).

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[4] 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[7,8], 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[3,16], 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[16], 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[17] and the serpin protease inhibitor involved in pathogen defence and programmed cell death[18].

Transposable elements were first reported in maize[19] and have since been shown to have important roles in shaping genome evolution and gene regulatory networks of many species[20]. Most of the maize

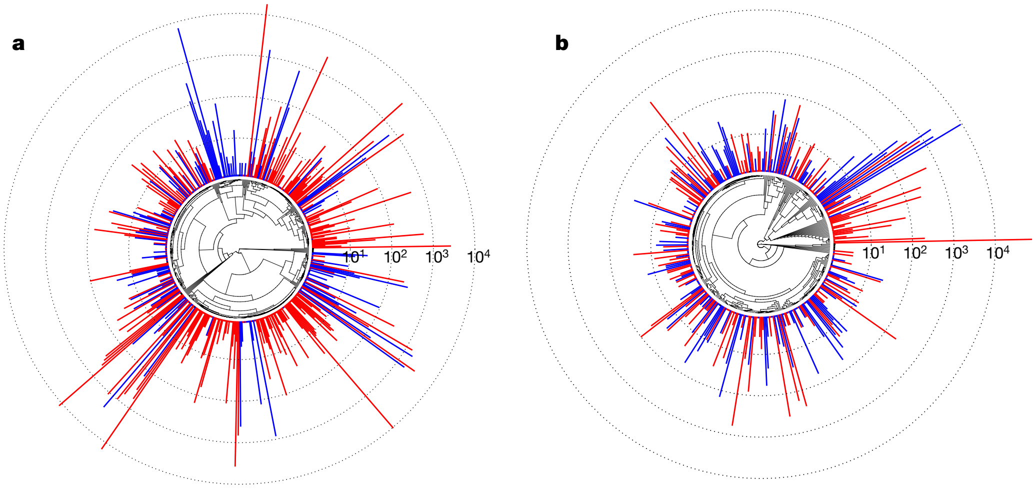


Figure 2 | Phylogeny of maize and sorghum LTR retrotransposon families. a, b, Both Ty3/Gypsy (**a**) and Ty1/Copia (**b**) superfamilies are present at higher copy number in maize (red) than in sorghum (blue). Bars ($\log_{10}$-scaled) depict family copy numbers.

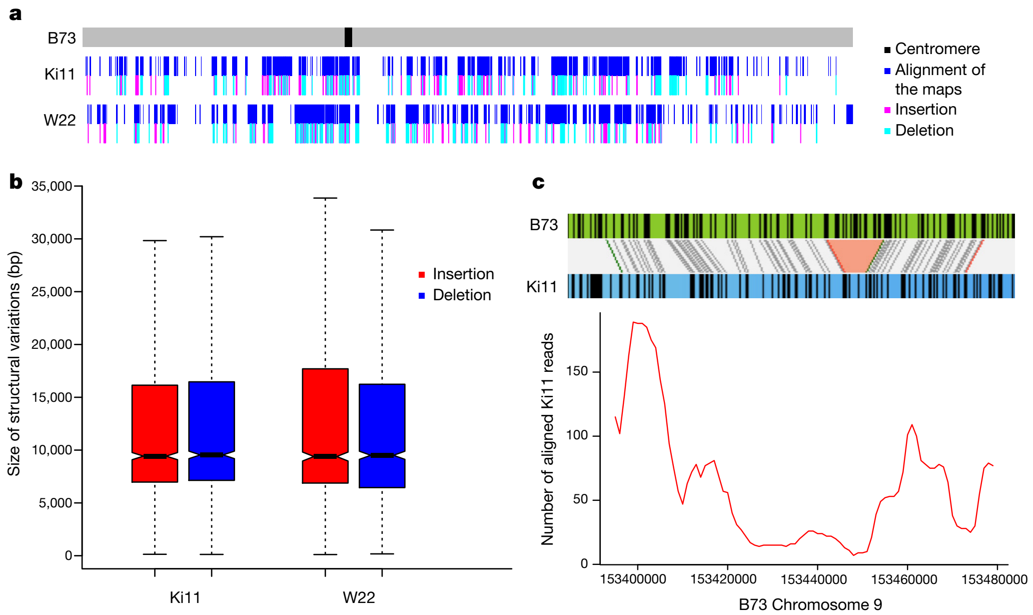


Figure 3 | Structural variation from Ki11 and W22. a, Alignment and structural variation called from Ki11 and W22 optical maps on chromosome 10. **b**, Size distribution of the insertion and deletions in Ki11 and W22. **c**, Example of using short-read alignment to verify a missing region mapped in Ki11.

genome is derived from transposable elements[3,21], and careful study of a few regions has revealed a characteristic structure of sequentially nested retrotransposons[21,22] and the effect of deletions and recombination on retrotransposon evolution[23]. In the annotation of the original maize assembly, however, fewer than 1% of long terminal repeat (LTR) retrotransposon copies were intact[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[25], and both nucleotide polymorphisms and structural variations have important roles in its phenotypic variation[10,26]. However, genome-wide patterns of structural variation in plant genomes are difficult to assess[27], and previous efforts have relied on short-read mapping, which misses the vast majority of intergenic spaces where most rearrangements are likely to occur[10]. To investigate structural variation at a genome-wide scale, we generated optical maps (Extended Data Table 1) for two additional maize inbred lines: the tropical line Ki11, one of the founders of the maize nested association mapping (NAM) population[28], and W22, which has served as a foundation for studies of maize genetics[29]. Owing to the high degree of genomic diversity among these lines, only 32% of the assembled 2,216 Mb map of Ki11 and 39% of the 2,280 Mb W22 map could be mapped to our new B73 reference via common restriction patterns (Table 2, Fig. 3a and Extended Data Fig. 7). The high density of alignments across and near many of the exceedingly retrotransposon-rich centromeres reflects the comparatively low genetic diversity of most centromeres in domesticated maize[15] and illustrates the ability of the combined optical mapping/single-molecule sequencing methodology to traverse large repeat-rich regions. Within the aligned regions, approximately 32% of the Ki11 and 26% of the W22 optical maps exhibited clear evidence of structural variation, including 3,408 insertions and 3,298 deletions (Table 2). The average indel size was approximately 20 kb, with a range from 1 kb to over 1 Mb (Fig. 3b). 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[30].

Although maize is often considered to be a large-genome crop, most major food crops have even larger genomes with more complex repeat landscapes[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. & Snowdon, R. J. Accessing complex crop genomes with next-generation sequencing. *Theor. Appl. Genet.* **126,** 1–11 (2013).
2. Morrell, 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).

7. Wallace, J. G. *et al.* Association mapping across numerous traits reveals patterns of functional variation in maize. *PLoS Genet.* **10,** e1004845 (2014).
8. Rodgers-Melnick, E., Vera, D. L., Bass, H. W. & Buckler, E. S. Open chromatin reveals the functional maize genome. *Proc. Natl Acad. Sci. USA* **113,** E3177–E3184 (2016).
9. Hufford, M. B. *et al.* Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **44,** 808–811 (2012).
10. Chia, J. M. *et al.* Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.* **44,** 803–807 (2012).
11. Wei, F. *et al.* The physical and genetic framework of the maize B73 genome. *PLoS Genet.* **5,** e1000715 (2009).
12. Ganal, M. W. *et al.* A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS One* **6,** e28334 (2011).
13. Hirsch, C. N. *et al.* Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. *Plant Cell* **28,** 2700–2714 (2016).
14. Gent, J. I., Wang, K., Jiang, J. & Dawe, R. K. Stable patterns of CENH3 occupancy through maize lineages containing genetically similar centromeres. *Genetics* **200,** 1105–1116 (2015).
15. Schneider, K. L., Xie, Z., Wolfgruber, T. K. & Presting, G. G. Inbreeding drives maize centromere evolution. *Proc. Natl Acad. Sci. USA* **113,** E987–E996 (2016).
16. Schnable, J. C., Springer, N. M. & Freeling, M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl Acad. Sci. USA* **108,** 4069–4074 (2011).
17. McHale, L., Tan, X., Koehl, P. & Michelmore, R. W. Plant NBS-LRR proteins: adaptable guards. *Genome Biol.* **7,** 212 (2006).
18. Fluhr, R., Lampl, N. & Roberts, T. H. Serpin protease inhibitors in plant biology. *Physiol. Plant.* **145,** 95–102 (2012).
19. McClintock, B. The origin and behavior of mutable loci in maize. *Proc. Natl Acad. Sci. USA* **36,** 344–355 (1950).
20. Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* **8,** 272–285 (2007).
21. SanMiguel, P. *et al.* Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274,** 765–768 (1996).
22. Brunner, S., Fengler, K., Morgante, M., Tingey, S. & Rafalski, A. Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell* **17,** 343–360 (2005).
23. Sharma, A., Schneider, K. L. & Presting, G. G. Sustained retrotransposition is mediated by nucleotide deletions and interelement recombinations. *Proc. Natl Acad. Sci. USA* **105,** 15470–15474 (2008).
24. Baucom, R. S. *et al.* Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genet.* **5,** e1000732 (2009).
25. Buckler, E. S., Gaut, B. S. & McMullen, M. D. Molecular and functional diversity of maize. *Curr. Opin. Plant Biol.* **9,** 172–176 (2006).
26. Dooner, H. K. & He, L. Maize genome structure variation: interplay between retrotransposon polymorphisms and genic recombination. *Plant Cell* **20,** 249–258 (2008).
27. Saxena, R. K., Edwards, D. & Varshney, R. K. Structural variations in plant genomes. *Brief. Funct. Genomics* **13,** 296–307 (2014).
28. McMullen, M. D. *et al.* Genetic properties of the maize nested association mapping population. *Science* **325,** 737–740 (2009).
29. Strable, J. & Scanlon, M. J. Maize (*Zea mays*): a model organism for basic and applied research in plant biology. *Cold Spring Harb. Protoc.* **2009,** pdb.emo132 (2009).
30. Swanson-Wagner, R. A. *et al.* Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res.* **20,** 1689–1699 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements Y.J., B.W., J.S., M.S.C., X.W., S.K. and D.W. were supported by NSF Gramene grant IOS-1127112, NSF Cereal Gene Discovery grant 1032105, USDA-ARS CRIS 1907-21000-030-00D and NSF Plant Genome award 1238014. J.R.-I. would like to acknowledge support from USDA Hatch project CA-D-PLS-2066-H and NSF Plant Genome award 1238014. R.K.D. would like to acknowledge support from NSF Plant Genome award 1444514. G.G.P. acknowledges support from NSF grant 1444624 and USDA NIFA project HAW05022-H. M.S.C would also like to acknowledge support from NSF PGRP PRFB 1523793. The authors thank S. Koren for sharing genome assembly-related scripts.

Author Contributions D.W. and Y.J. designed and conceived the research, M.M and K.G. prepared DNA samples for PacBio SMRT sequencing, D.R.R., P.P., E.A. and W.R.M. performed PacBio SMRT sequencing, B.W., J.S., R.K.D., T.L. and A.H. generated the BioNano optical genome maps, M.R. generated Illumina sequencing data, Y.J., T.L., J.S., C.-S.C. and A.H. performed the genome assembly, J.C.S., M.S.C., X.W., B.W., Y.J. and S.K. performed gene annotation and evolutionary studies, M.C.S., M.R.M., N.M.S. and J.R.-I. performed transposable element analysis, J.G., J.S., R.K.D., K.L.S., T.K.W., G.G.P. and Y.J. performed the analysis of centromeres and telomeres. B.W., Y.J., J.S., T.L., A.H. and R.K.D. performed the structural variation study. X.W., J.C.S. and Y.J. contributed to the data release. Y.J., J.R.-I., R.K.D., G.G.P. and D.W. wrote the paper. All authors contributed to the revision of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to D.W. (ware@cshl.edu).

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 inbred line B73 from NCRPIS (PI550473), grown at University of Missouri. Seeds of this line were deposited at NCRPIS (tracking number PI677128). Etiolated seedlings were grown for 4–6 days in Pro-Mix at 37 °C in darkness 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 Megarupter 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 described[31]. DNA was labelled at Nt.BspQI sites using the IrysPrep kit. Molecules collected from BioNano chips were *de novo* assembled as previously described[32] 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 PBcR –MHAP pipeline[33] and Falcon[34] 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 system[35]. 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 RefAligner[32] ($P < 1 \times 10^{-11}$). 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 package[36] with the following parameter settings: '-l(minimum length of a single match) 100 -c(the minimum length of a cluster of matches) 1000'. To only use unique hits as markers, alignment hits were filtered with the following parameters: '-i(the minimum alignment identity) 98 -l(the 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 from an intermated maize recombinant inbred line population (Mo17 × B73)[10]. Contigs with markers not located in pseudochromosomes from the physical 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 Pbjelly (-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-mem[37]. Sequence error correction was performed with the Pilon pipeline[38], after aligning reads with BWA-mem[37] and parsing with SAMtools[39], 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 LTRharvest[40], LTRdigest[41], SINE-Finder[42], MGEScan-non-LTR[43], MITE-hunter[44], HelitronScanner[45], and others (details in Supplementary Information). The scripts, parameters, and intermediate files of each transposable element superfamily are available at https://github.com/mcstitzer/maize_v4_TE_annotation.

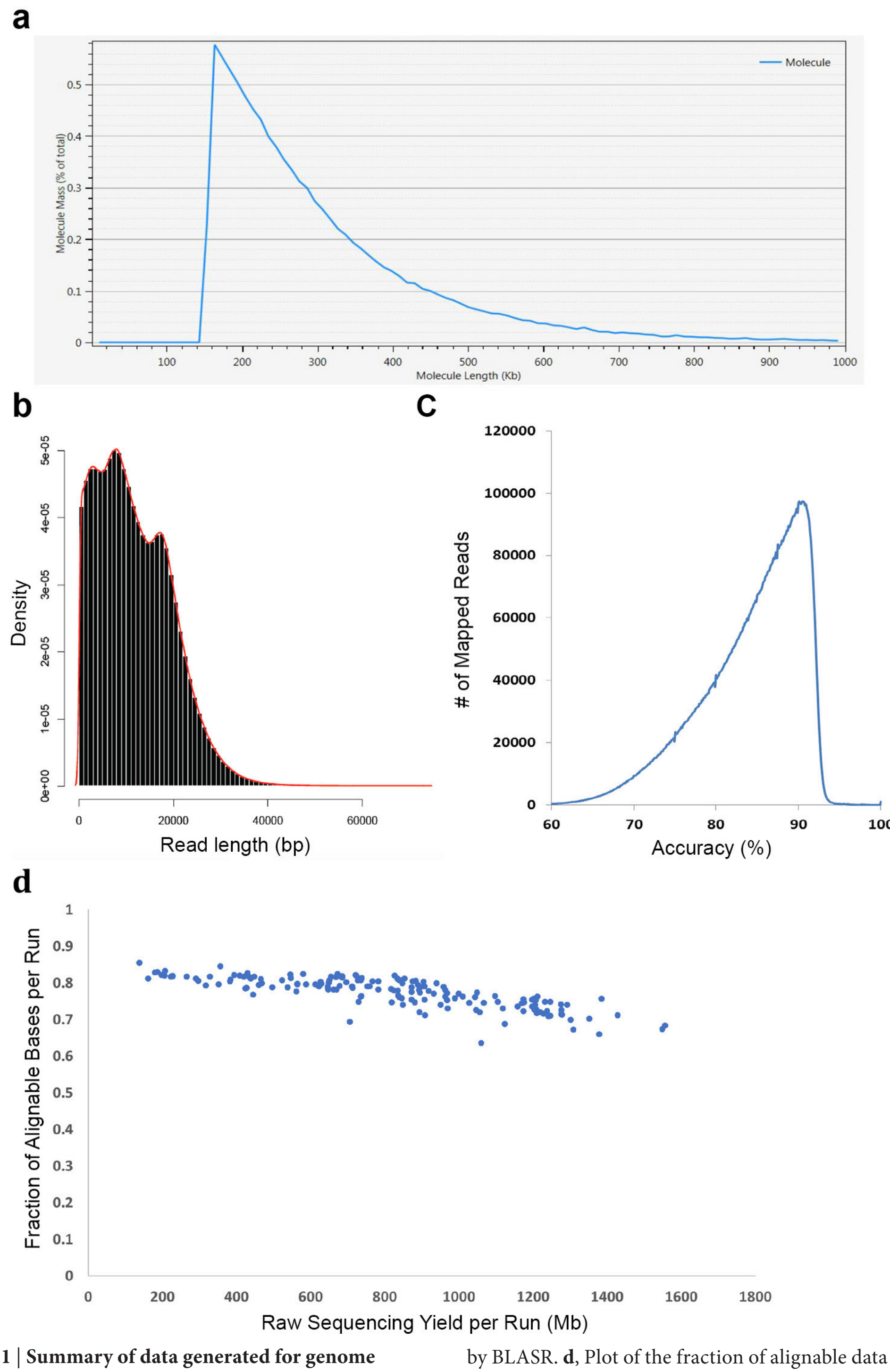
The MAKER-P pipeline was used to annotate protein-coding genes[46], integrating *ab initio* prediction with publicly available evidence from full-length cDNA[47], *de novo* assembled transcripts from short-read mRNA sequencing (mRNA-seq)[48], isoform-sequencing (Iso-Seq) full-length transcripts[14], 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.

Structural variation. Leaves were used to prepare high molecular mass DNA and optical genome maps were constructed as described above for B73. Structural variant calls were generated based on alignment to the reference map B73 v4 chromosomal assembly using the multiple local alignment algorithm (RefSplit)[32]. A structural variant was identified as an alignment outlier[32,49], defined as two well-aligned regions separated by a poorly aligned region with a large size difference between the reference genome and the map or by one or more unaligned sites, or alternatively as a gap between two local alignments. A confidence score was generated by comparing the non-normalized P values of the two well-aligned regions and the non-normalized log-likelihood ratio[50] of the unaligned or poorly aligned region. With a confidence score threshold of 3, RefSplit is sensitive to insertions and deletions as small as 100 bp (events smaller than 1 kb are generally compound or substitution and include label changes, not just spacing differences) and other changes such as inversions and complex events which could be balanced. Insertion and deletion calls were based on an alignment outlier P-value threshold of 1×10^{-4}. Insertions or deletions that crossed gaps in the B73 pseudomolecules, or that were heterozygous in the optical genome maps, were excluded. Considering the resolution of the BioNano optical map, only insertion and deletions larger than 100 bp were used for subsequent analyses. To obtain high-confidence deletion sequences, sequencing reads from the maize HapMap2 project[8] for Ki11 and W22 were aligned to our new B73 v4 reference genome using Bowtie2 (ref. 51). Read depth (minimum mapping quality >20) was calculated in 10-kb windows with step size of 1 kb. Windows with read depth below 10 in Ki11 and 20 in W22 (sequencing depths for Ki11 and W22 were 2.32× and 4.04×, respectively) in the deleted region were retained for further analysis.

Data availability. Raw reads, genome assembly sequences, and gene annotations 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 LPUQ00000000. 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.

31. VanBuren, R. *et al.* Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature* **527,** 508–511 (2015).
32. Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12,** 780–786 (2015).
33. Berlin, K. *et al.* Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33,** 623–630 (2015).
34. Chin, C. S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13,** 1050–1054 (2016).
35. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **44,** D67–D72 (2016).
36. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5,** R12 (2004).
37. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
38. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9,** e112963 (2014).
39. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
40. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9,** 18 (2008).
41. 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).
42. Wenke, 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).
43. Rho, M. & Tang, H. MGEScan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes. *Nucleic Acids Res.* **37,** e143 (2009).
44. Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **38,** e199 (2010).

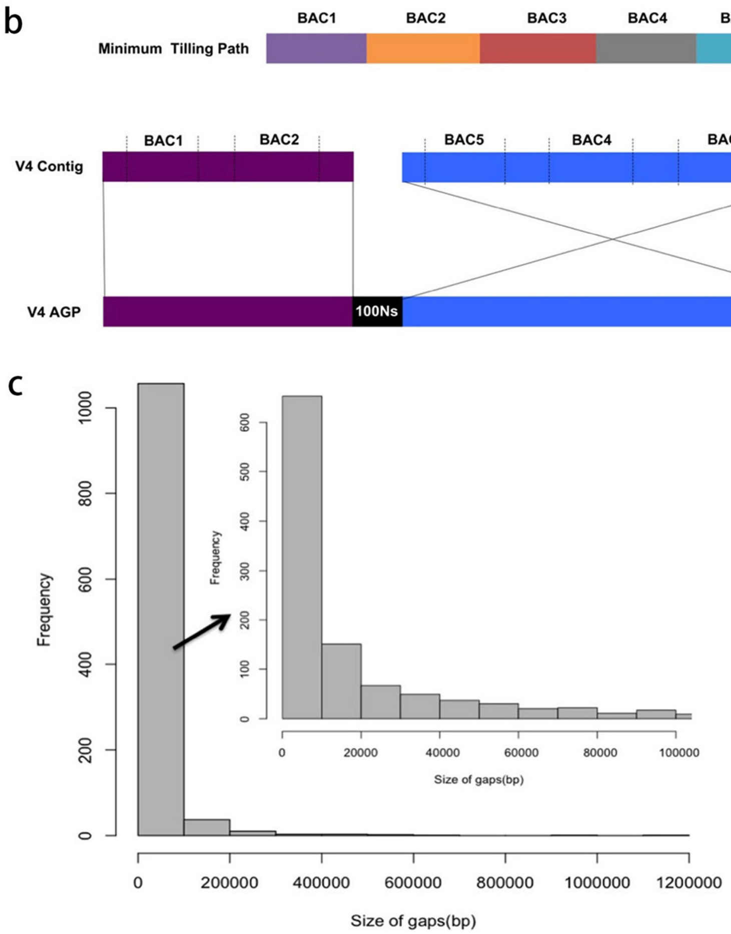
45. Xiong, W., He, L., Lai, J., Dooner, H. K. & Du, C. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proc. Natl Acad. Sci. USA* **111,** 10263–10268 (2014).
46. Campbell, M. S. *et al.* MAKER-P: a tool kit for the rapid creation, management, and quality 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).



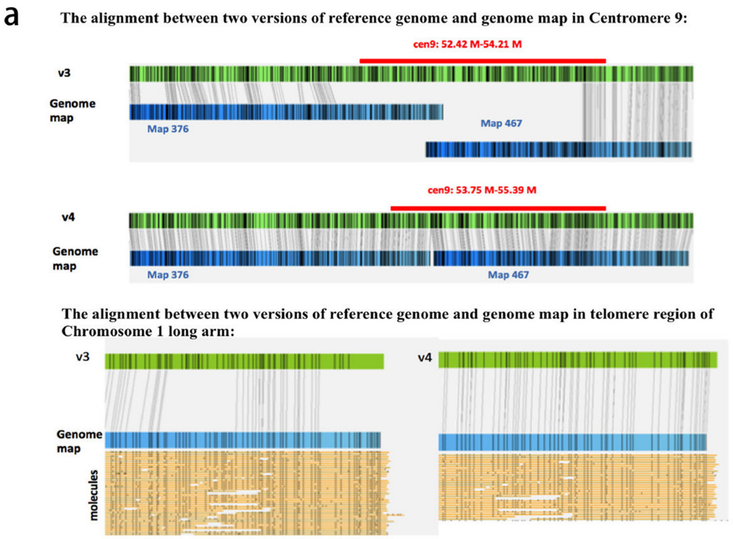
Extended Data Figure 1 | Summary of data generated for genome construction. a, Size distribution of single molecules for the optical maps. A total of 150 Gb (~60-fold coverage) of single-molecule raw data from BioNano chips was collected for map construction. The N50 of the single molecules was ~261 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.

a

	No. of contigs	N50 (bp)	No. of contigs above N50 size	Max contig size (bp)	Assembly size (Gb)	Conflicts with genome map
Falcon	4,845	1,746,430	391	6,555,927	2.15	704
PBcR-MHAP (k=16)	7,729	380,973	1647	2,234,976	2.08	72
PBcR-MHAP (k=14)	3,303	1,038,844	615	5,651,342	2.10	36



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.



The alignment between two versions of reference genome and genome map in Centromere 9:

The alignment between two versions of reference genome and genome map in telomere region of Chromosome 1 long arm:

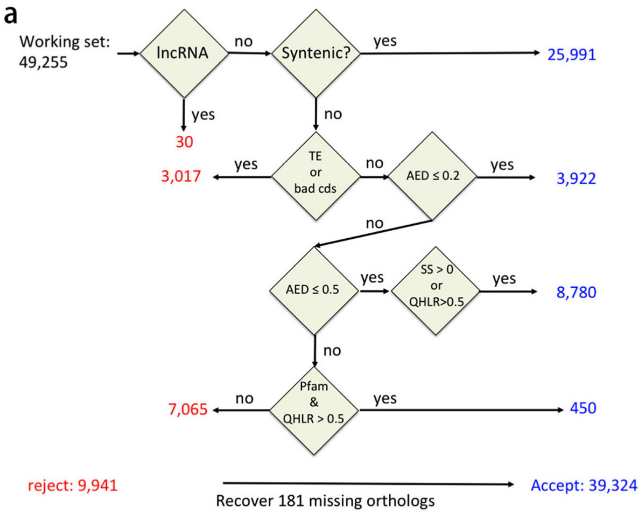
b

V4				V3			
Chr.	Start	End	Size (Mb)	Chr.	Start	End	Size (Mb)
1	136.77	137.12	0.35	1	134.22	134.96	0.74
2	95.51	97.49	1.98	2	93.52	95.37	1.85
3	85.78	86.93	1.15	3	85.34	85.63	0.3
				3	99.79	101	1.21
4	109.07	110.5	1.43	4	103.6	104.02	0.42
				4	105.36	106.21	0.85
5	104.54	106.82	2.28	5	101.98	104.19	2.21
6	52.3	53.11	0.8	6	39.15	39.31	0.16
				6	49.75	50.37	0.62
7	56.38	56.68	0.3	7	22.73	23.23	0.5
				7	54.62	54.89	0.27
				7	60.41	60.59	0.18
				7	62.34	62.44	0.1
8	50.53	52.07	1.54	8	48.15	48.31	0.16
				8	49.06	50.97	1.92
9	53.75	55.39	1.65	9	52.42	54.21	1.79
9	57.36	57.76	0.4				
10	51.39	52.78	1.39	10	0	0.51	0.51
				10	50.07	51.81	1.74
				scaffold_498	0.17	0.46	0.29
				scaffold_507	0.72	0.85	0.12

c

Chromosome arm	Telomere Repeats?	Number of telomere repeats in the first or last 5kb	Other repeats
1S	no	NA	142 copies of Knob180 in the first 50 kb
1L	yes	397	
2S	no	NA	Continuous (ACT) simple repeat in the first 3 kb
2L	no	NA	
3S	no	NA	117 copies of knob180 in the first 50 kb
3L	no	NA	
4S	no	NA	
4L	yes	88	
5S	no	NA	
5L	yes	486	
6S	no	NA	77 copies of knob180 in the first 50 kb
6L	yes	322	
7S	yes	341	
7L	yes	420	
8S	no	NA	116 copies of knob180 in the first 50 kb
8L	yes	96	
9S	no	NA	58 copies of knob180 in the first 50 kb
9L	no	NA	
10S	yes	360	
10L	yes	409	

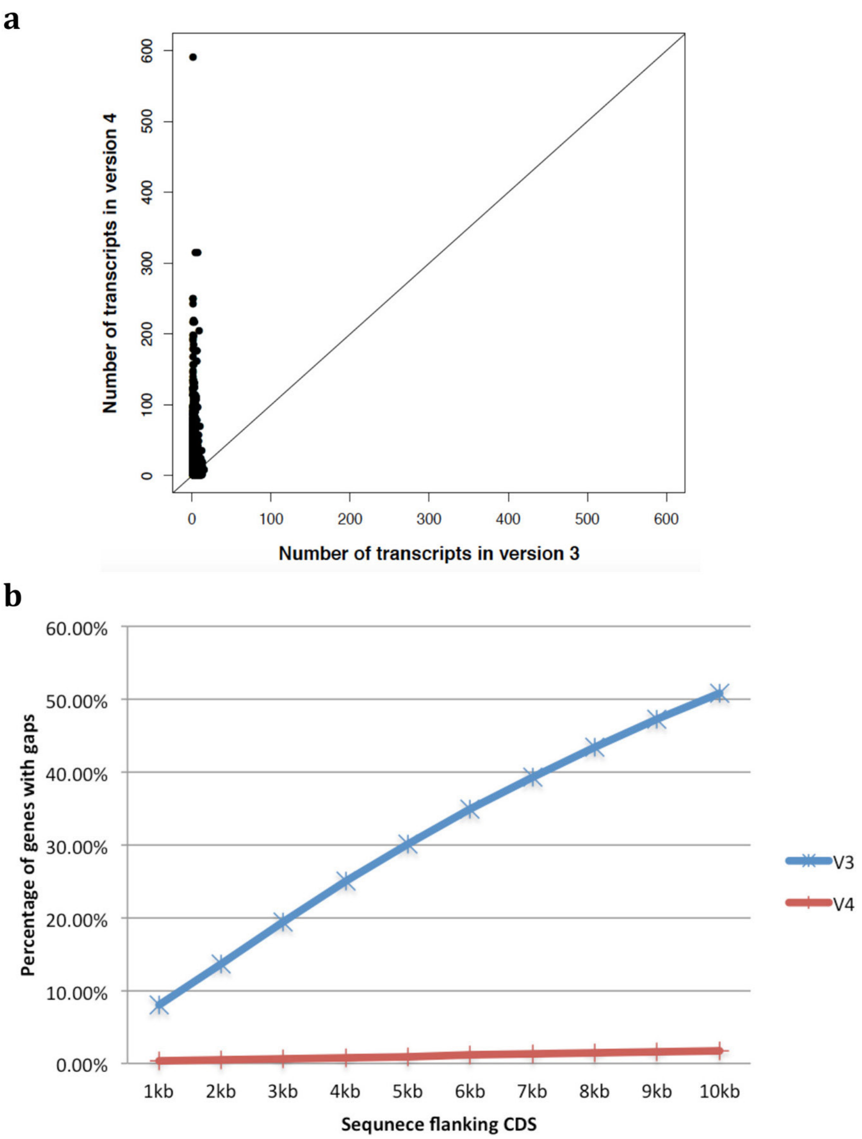
Extended Data Figure 3 | Quality assessment and comparison of the assembly in centromere and telomere regions in maize B73 RefGen_v3 and v4. **a**, Quality assessment of centromere and telomere using optical genome map. **b**, Locations of centromeres on pseudomolecules defined by ChIP–seq in the B73 RefGen_v3 and v4. **c**, Telomere repeats found in the B73 RefGen_v4 pseudomolecules.



b

	V4 Filtered set	V3 Filtered set (anchored on chromosomes)
Number of protein-coding genes	39,324	39,323
Number of transcripts	131,319	63,074
Average number of transcripts per gene	3.3	1.6
Median exons per gene (based on transcript with the longest CDS)	4	3
Median exon length (based on transcript with the longest CDS)	156	159
Median transcript length (based on transcript with the longest CDS)	1,281	1,374
Median CDS length (based on transcript with the longest CDS)	951	924
Genes with annotated 5’ UTR	26,035	29,705
Genes with annotated 3’ UTR	25,383	30,228
Transcripts with annotated 5’ UTR	112,660	51,419
Transcripts with annotated 3’ UTR	113,637	52,080

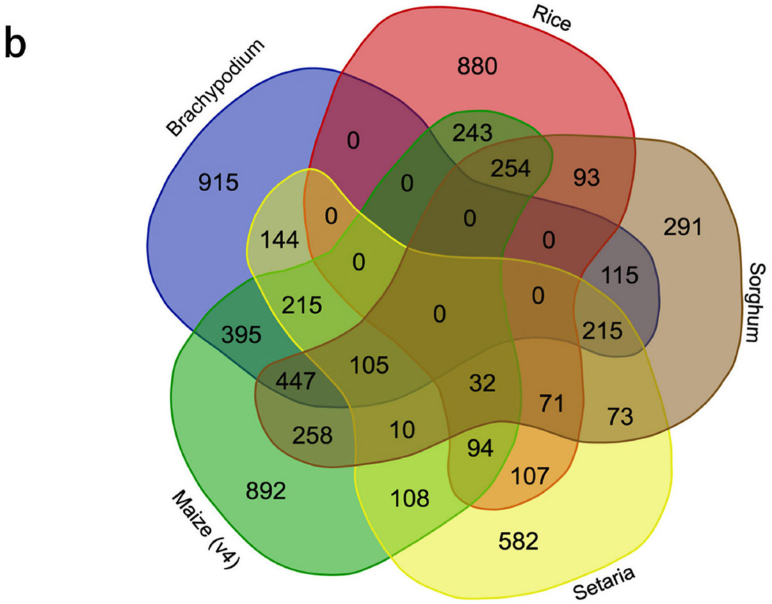
Extended Data Figure 4 | Details of the gene annotation of maize B73 RefGen_v4. a, The pipeline used to characterize high confidence gene models. **b**, Summary of B73 RefGen_v4 protein-coding gene annotation, and comparison with RefGen_v3 annotation.



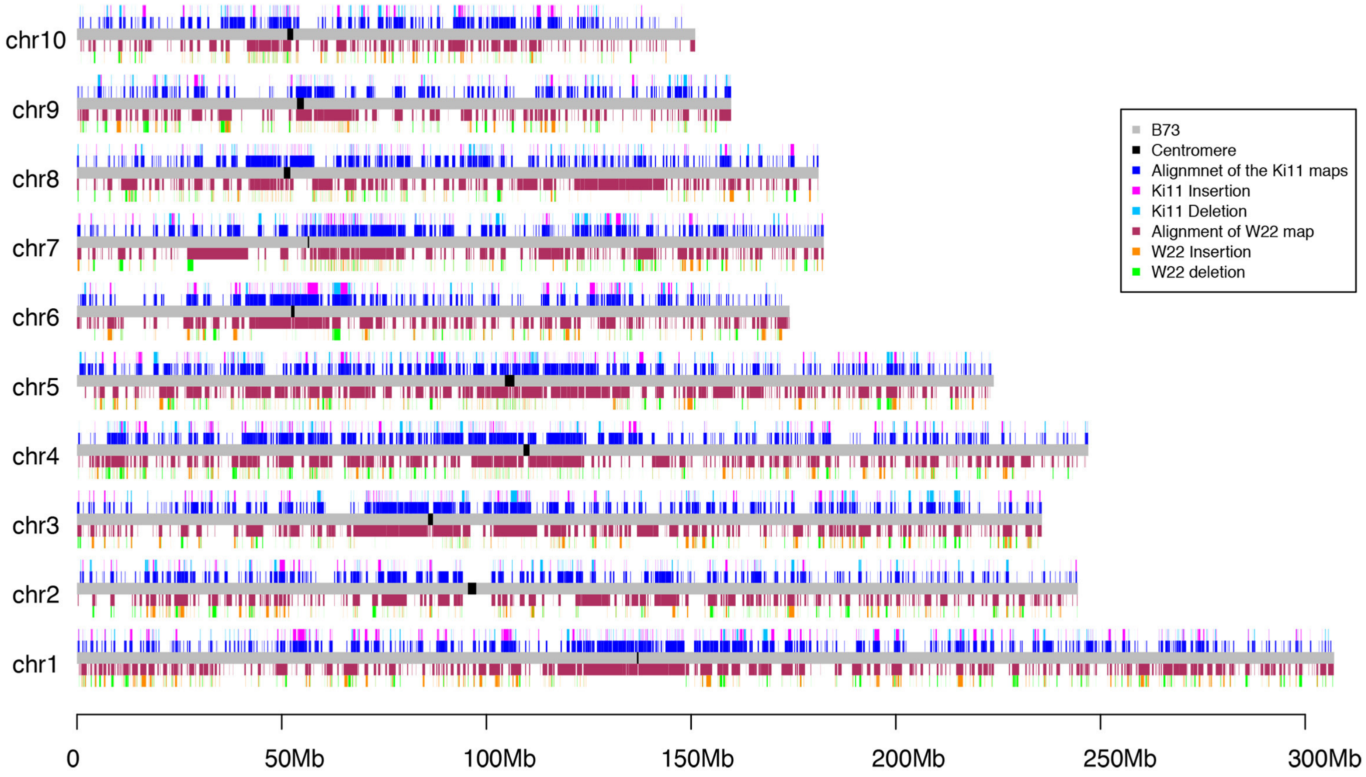
Extended Data Figure 5 | Improvement of the annotation of alternative splicing and completeness of regulatory regions of maize RefGen_v4 genes. **a**, Number of transcripts of each gene in v3 and v4 annotation. **b**, Percentages of genes with gaps in flanking regions in the v3 and v4 annotations.

a

Species	Clade of most recent common ancestor				
	Poaceae	*Panicoideae*	*Andropogoneae*	*Zea*	*BEP*
Zea mays (v4)	18,995 (86.2)	733 (58.7)	238 (81.0)	1,925 (100.0)	na
Zea mays (v3)	18,854 (85.5)	717 (57.5)	250 (85.0)	1,925 (100.0)	na
Sorghum bicolor	20,084 (91.1)	943 (75.6)	294 (100.0)	na	na
Setaria italica	20,292 (92.0)	1,248 (100.0)	na	na	na
Oryza sativa	20,274 (92.0)	na	na	na	419 (100.0)
Brachypodium distachyon	19,497 (88.4)	na	na	na	419 (100.0)
Total	22,048 (68.7)	1,248 (36.5)	294 (66.0)	1,925 (100.0)	419 (100.0)



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		B73	Ki11	W22
10–500 kb	# of Maps	311	675	540
	Quantity (Mb)	102.462	213.642	179.105
	Bin proportion (% by mass)	5%	10%	7%
500–1000 kb	# of Maps	323	644	710
	Quantity (Mb)	237.117	465.86	526.351
	Bin proportion (% by mass)	11%	21%	21%
1000–2000 kb	# of Maps	341	573	606
	Quantity (Mb)	486.497	805.219	850.356
	Bin proportion (% by mass)	23%	36%	34%
>2000 kb	# of Maps	378	256	331
	Quantity (Mb)	1293.607	731.371	974.819
	Bin proportion (% by mass)	61%	33%	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%