The complex sequence landscape of maize revealed by single molecule technologies

1. Data generation

Whole-genome sequencing using SMRT (Single Molecule Real-Time) technology: DNA samples for SMRT sequencing were prepared using maize inbred line B73 from NCRPIS (PI550473), which was grown at University of Missouri. The seeds used for sequencing were deposited at NCRPIS (tracking number: PI 677128). Kernels were placed in a flat with Pro-Mix and allowed to grow for 4–6 days in the dark at 37°C. To eliminate chloroplast DNA, etiolated tissue was harvested. 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).

Shearing of genomic DNA to a size range of 15–40 kb was performed using either G-tubes (Covaris®) or a Megarupter® device (Diagenode). Sheared DNA was enzymatically repaired and converted into SMRTbell™ template libraries as recommended by Pacific Biosciences. Briefly, hairpin adapters were ligated to the sheared and repaired DNA fragments. Damaged DNA fragments and those without hairpin adapters ligated at both ends were eliminated by digestion with Exonuclease III and Exonuclease VII (New England BioLabs). The resulting SMRTbell template libraries were subjected to Blue Pippin preparative electrophoresis purification (Sage Sciences) to select insert sizes ranging from 15 to 50 kb. These size-selected SMRTbell template libraries were used in subsequent sequencing steps. Sequencing was performed on a PacBio RS II instrument. All sequencing runs were performed with P6-C4 sequencing chemistry. To acquire long reads, all data were collected as either 5- or 6-hour sequencing movies.

Construction of genome maps using the Irys system: High–molecular weight genomic DNA was isolated from maize line B73 (PI550473). Three grams of fresh young ear tissue was fixed with 2% formaldehyde. After blending in isolation buffer using a tissue homogenizer, the sample was filtered and washed in Triton X-100–containing buffer as previously described[34]. Nuclei were purified on Percoll cushions, washed extensively, and embedded in low-melting point agarose at various dilutions. Finally, the DNA plugs were treated with lysis buffer containing detergent, proteinase K, and β-mercaptoethanol (BME). DNA was labeled using Nt.BspQI and stained based on the IrysPrep kit and protocol. Molecules collected from BioNano chips were *de novo* assembled as previously described in previous study[35] using "optArgument_human".

2. Genome assembly

***De novo* assembly of the long reads**: Two assembly tools, PBcR-MHAP and FALCON, were independently evaluated for *de novo* assembly of PacBio SMRT Sequencing reads. For PBcR, following the recommended parameters for large genome assembly[1], *k*-mer lengths of 16 and 14 were used to test the performance of assembler. Unitigs were filtered according to sequencing coverage, according to the following criteria: coverage ≥ 2 reads, and a single read must not cross more 50% of a contig. FALCON v.0.4 (https://github.com/PacificBiosciences/FALCON-integrate/tree/0.4.0) was also used for *de novo* assembly. The overall design of FALCON follows the hierarchical genome-assembly process[2]. Instead of BLASR, daligner was used to overlap reads. To lay out contigs from the assembly graph, the error-correction module was updated, and the Celera Assembler was replaced by a string graph–based module. Due to the highly repetitive nature of the maize genome, we adopted more aggressive parameters to reduce computation time. For the full data set, only reads longer than 12 kb were corrected. To identify overlaps between raw sequences, we used "-M24 -l4800 -k18 -h480 -w8" for Daligner. Using these parameters, only overlaps longer than 4,800 bp were considered for error correction with seed matches > 480 bp.

To ascertain the quality of the three independent assemblies (FALCON, PBcR with k=16 and 14), the BioNano scaffolding pipeline NGM Hybrid Scaffold (NGM-HS) (version 4304) was used to generate an *in silico* map for sequence contigs from each assembly. The maps were aligned against the genome assemblies using RefAligner[3,4] to identify and resolve potential conflicts in sequence contigs or genome maps. The result showed that the PBcR–MHAP assembly (k=14) had the fewest conflicts (Extended Data Figure); consequently, it was adopted as the new B73 genome reference.

Curation of the assembly: Comparison between the contigs and genome map identified 36 conflicts. Next Generation Mapping (NGM-HS) from BioNano Genomics' Irys® System was used to resolve conflicts between the sequence and genome map assemblies by cutting either assembly (option: –N 2 –B 2); cut decisions were based on chimeric scores of labels near the conflict junctions on the genome map. The chimeric score of a label represents the percentage of BioNano molecules that can fully align to the genome map 55 kb to the left or right of that label. If the chimeric scores of all labels within 10 kb of the conflict junction were ≥ 35, the scaffolding pipeline suggested a cut in the sequence contig. If any label in the region had a chimeric score < 35, a cut was suggested in the BioNano genome map. All proposed cuts were manually evaluated using BioNano molecule-to-genome map alignments, molecule-to-sequence contig alignments, and the BAC-based fingerprint map. Of these 36 conflicts, 18 were chimeric in the long reads assembly, and 13 were chimeric in the BioNano map; five were left unresolved.

Using alignments of the genome map, a total of 1,369 overlaps were detected among the tails of the contigs. There are two possible reasons for this: the overlaps could be repeat boundaries between contigs from the Celera assembler[5], or alternatively, highly identical repeats could be over-collapsed in the BioNano map. The redundancy at the edges of nearby contigs generated by the Celera assembler was resolved as follows: if two contigs were detected to have overlap from 0.5-10 kb (based on the size of PacBio reads) by BioNano genome map and had sequence identity over 95% in the overlapped region, they were merged by Mininus2[6]. A total of 670 contigs were merged into 310 larger contigs.

Pseudomolecule construction: Using unique BAC sequences as markers, we could order and orient 315 super-scaffolds and 25 contigs. In addition, we also incorporated a genetic map built from an intermated maize recombinant inbred line population (Mo17 × B73)[7] to complement pseudomolecule construction and validation. In this new AGP (A Golden Path) of the maize reference genome, a total of 331 hybrid scaffolds and 45 contigs were ordered and oriented. Of 1,907 markers on the genetic map, 1,868 could be mapped to the new pseudomolecules, with only one disagreement, demonstrating the high accuracy of the AGP. During the following gap-filling procedure, 170 gaps were filled by SMRT long reads. To ensure base-pairing accuracy and further polish the pseudomolecules, we deployed Illumina reads with 100-fold coverage of the genome; as a result, about 80,000 bases were corrected as single-nucleotide changes or small insertions or deletions.

Centromere identification by ChIP-seq: Peaks of CENH3 enrichment were defined by CENH3 ChIP-seq as described previously[8] using the HOMER findPeaks software[9]. Input reads from the CENH3 ChIP sample were used as controls. All reads were mapped to the genome using BWA-MEM[10]. As a first step, all reads, including potential repetitively mapping reads, were used to identify a set of putative CENH3-enriched regions; the parameters of HOMER findPeaks were set as follows: -region -size 5000 –minDist 50000 -F 8 -L 0 -C -0. A set of high-confidence peaks was then independently identified using only uniquely mapping reads (as defined by MAPQ values ≥ 20) with the following parameters: -region -size 5000 -F 16 -L 0 -C -0. Putative CENH3-enriched regions that were either shorter than 100 kb, or that did not overlap with at least one high-confidence peak, were discarded. To generate the final set of centromeric loci, the remaining putative CENH3-enriched regions were merged if they were less than 500 kb apart.

3. Comparison of genome assembly quality in Maize B73 RefGen_v3 and v4

The Maize Genome Sequencing Pilot Project randomly selected 100 BAC clones for high-quality sequencing, resulting in 98 curated BACs of finished quality[11]. In total, 25 of the 28 fully completed BACs were spanned by a single contig in RefGen_v4, with identity above 99.9%. In addition, the maize pilot sequence contains 57 BACs with ordered contigs and gaps. The gaps of 46 BACs could be closed by a single contig in RefGen_v4.

Several gene models with assembly errors in the maize B73 RefGen_v3 have been corrected in the current maize genome. For example, the *rgh3* locus (JN692485.1) was involved in an assembly error that arose due to incorrect ordering and orientation of contigs in the BAC sequence, resulting in mis-annotation of this gene as two distinct gene models[12]. This problem was successfully fixed in the v4 assembly. Due to correction of such errors and the increase in contiguity described above, the RefGen_v4 assembly is much more robust as a reference genome than the old BAC sequences.

4. Gene annotation

Generation of a working gene set: MAKER-P version 3.1[13] was used to annotate the maize RefGen_v4 genome. As evidence, we used all annotated proteins from *Sorghum bicolor*, *Oryza sativa*, *Setaria italica*, *Brachypodium distachyon*, and *Arabidopsis thaliana*, downloaded from Gramene.org release 48[14]. For transcript evidence, the 111,151 high quality transcripts from Iso-seq were further polished by illumina RNA-seq reads generated from same tissues[15] using Ectools (https://github.com/jgurtowski/ectools). Another set of 69,163 publicly available full-length cDNAs deposited in Genbank[16], 1,574,442 Trinity-assembled transcripts from 94 B73 RNA-Seq experiments[17], and 112,963 transcripts assembled from deep sequencing of a B73 seedling[18] were also included as transcript evidence. For gene prediction, we used Augustus[19] and FGENESH (http://www.softberry.com/berry.phtml) trained on maize and monocots, respectively. For repeat masking, we used RepeatMasker and the B73-specific TE exemplars[20]. Helitron elements and captured exons within pack-MULES were removed from this library to prevent the masking of non–TE-related protein-coding genes. Additional masking was performed using a set of known TE-derived proteins distributed with the MAKER software package[13].

The final annotation set was built iteratively. The first step, which included all of the protein evidence, the full-length cDNAs from GenBank, and the Iso-Seq data, generated 34,088 genes with 56,671 transcripts. For the second step, the gene models from the first pass were given back to MAKER as models, allowing them to persist unchanged in the annotation set. Next, the additional transcript evidence derived from short reads was included. This step generated an additional 9,548 genes with 11,475 transcripts. To retain as many genes as possible from the v3 annotations, the third pass added the previously annotated B73 transcripts and protein translations from the v3 assembly as evidence. This step added 5,449 genes with 5,947 transcripts. MAKER-P is conservative in annotating alternate transcripts. Additionally, transcripts that contain large intron retentions, non-canonical splicing, or are expressed at low levels are also difficult to annotate confidently by computational methods. However, the single-molecule Iso-Seq transcript sequencing method can unambiguously identify these hard-to-annotate transcripts. By including the additional unique Iso-Seq transcripts into the gene models from step 3, we generated a protein-coding gene annotation set of 49,085 genes and 161,680 transcripts (referred to as the working set).

Compara gene tree construction: The Ensembl Compara gene tree pipeline[21] was used to define gene families, construct phylogenetic gene trees, and infer orthologs and paralogs. Updated protocols used in the Ensembl version 81 software are detailed elsewhere (http://jul2015.archive.ensembl.org/info/genome/compara/homology_method.html). The analysis included annotated protein-coding genes from both the v3 and v4 gene sets of maize B73, as well as 17 additional species (five monocots, four dicots, one basal angiosperm, three lower plants, and four non-plants), which were downloaded from the Ensembl core databases within Gramene Release-41. Tree reconciliation to classify duplication and speciation nodes, and the assignment of taxon levels to nodes, used the following input species tree derived from the NCBI Taxonomy database[21]: ((((((((((sorghum_bicolor,(zea_mays_v3,zea_mays_v4)N)Andropogoneae,setaria_italica) Panicoideae,(brachypodium_distachyon,oryza_sativa)BEP_clade)Poaceae,musa_acuminata)commelinids,((((arabidopsis_thaliana,glycine_max),vitis_vinifera))rosids,solanum_lycopersicum)Eudicot)Mesangiospermae,amborella_trichopoda)Magnoliophyta,selaginella_moellendorffii)Tracheophyta,physcomitrella_patens)Embryophyta,chlamydomonas_reinhardtii)Viridiplantae,(((caenorhabditis_elegans,drosophila_melanogaster)Ecdysozoa,homo_sapiens)Bilateria,saccharomyces_cerevisiae)Opisthokonta)Eukaryota;.

Synteny maps, which relate collinear chains of orthologous genes between two genomes, were built using DAGchainer[22] in combination with other previously described methods[20,23].

Generation of the filtered gene set: The working set of protein-coding gene annotations is expected to contain TEs that were not masked prior to annotation, long noncoding RNAs annotated as protein-coding genes, and annotations with little supporting evidence. We filtered the working set based on evidentiary support, transposon screening, long noncoding RNA screening, homology support, and valid CDSs. The approach is schematically represented in Extended data Figure 4a.

tRNA annotation: tRNAs were identified using tRNAscan-SE[24] within the MAKER-P framework[25]. A total of 2,305 tRNAs were identified: 1,451 decode standard amino acids, four decode seleno-Cys, seven are putative suppressors, 13 contain an undeterminable anti-codon sequence, and 830 are apparent pseudogenes. Compared to the v3 assembly, v4 contains 59 additional complete tRNAs and 54 additional putative tRNA pseudogenes. This increase in identifiable tRNAs provides further evidence that v4 is a more complete genome assembly than v3.

5. Comparison of gene annotation between RefGen_v3 and v4

Alignment of v3 genes to the v4 genome: We used two pipelines to map the v3 genes to the v4 genome, Genome Assembly Converter and Mummer pipeline[26]. In Genome Assembly Converter, the ATAC pipeline[27] was used to create the alignment chain file between two assemblies, and then CrossMap[28] was used to convert the coordinates of the v3 gene annotation. A chromosome-to-chromosome alignment was first performed using Mummer to map the v3 genes to the v4 genome. Genes from v3 that could not be mapped to the same chromosome in v4 were then aligned to the whole v4 assembly. Only unique hits with identity above 98% and 100% coverage were retained for merging with the Genome Assembly Converter pipeline. Disagreements between the two pipelines were resolved as follows: if the Genome Assembly Converter pipeline had 100% coverage for a given gene, then those coordinates were kept; otherwise, the result from the Mummer alignment was used.

Alignment of the RefGen_v3 and v4 genome assemblies indicated that the two versions are highly consistent with each other in gene space. A total of 36,725 (94%) v3 gene models could be mapped to the new RefGen_v4 genome without sequence changes. Most of the remaining v3 genes (1,356) could be mapped, but crossed multiple contigs in RefGen_v3, with gaps; consequently, it is very likely that they were incorrectly assembled in v3. In RefGen_v4, most of these genes were contained within continuous sequences, indicating the improvement of the genomic sequences of these genes. In addition, 92 of the 146 genes previously unanchored in RefGen_v3 were anchored to chromosomes in the RefGen_v4 assembly.

Core promoter elements: Core promoter elements were analyzed in both RefGen_v3 and v4 with a published pipeline[29,30]. Comparison of core promoter elements, especially the TATA-box, CCAAT-box, and Y patch in the new assembly to those in the previously published assembly revealed 17.5% of genes in the new assembly contained a TATA-box, whereas in the previous assembly only 12.8% genes contained this element. Similarly, 7.2% genes contained a CCAAT-box and 58.17% contained Y patch in maize B73 RefGen_v4, versus 2.4% and 41.5%, respectively in v3.

Gene orientation: Of 30,926 genes that could be mapped between the v3 and v4 annotations，2,151 genes were switched to a different strand. To evaluate this, we compared gene orientation to sorghum orthologs within syntenic blocks. Among 652 genes that could be tracked in this manner, the orientation of 589 (90.3%) was conserved with sorghum (see the table below). Thus, in the vast majority of cases, the re-orientation of a gene in v4 brought the configuration into closer agreement with sorghum, further lending confidence to strand reassignments of v4 genes.

Identification of missing genes in maize genome: We identified 22,048 orthologous gene sets that originated prior to, or within, the grass common ancestor, and cataloged deficiencies in gene content among annotations of the five grass species (maize, rice, sorghum, *Setaria*, and *Brachypodium*). Of these sets, ~69% were found in all five species, and of individual species, rice, *Setaria*, and sorghum had the most complete representation, possessing from 91% to 92% of ortholog sets. By contrast, despite the fact that maize is a product of whole-genome duplication, maize genes were found in only 86% of ortholog sets, representing a deficit of over 3,000 genes. To minimize artifacts, we restricted analysis to 592 ortholog sets containing 668 sorghum genes that 1) are syntenic with an outgroup species (either rice, *Brachypodium*, or *Setaria*), 2) are flanked by genes contained within a synteny block that maps to a single maize contig in both the A and B subgenomes, and 3) lack alignment of CDS features to the v4 reference.

5. Structural identification of transposable elements

LTR retrotransposons: LTR retrotransposons were identified using LTRharvest[31] and LTRdigest[32]. LTRharvest searches sequence data for structural characteristics of LTR retrotransposons; in an analysis of the *Drosophila* X chromosome, it was shown to be the most sensitive among available structural search tools[33]. To be consistent with known LTR retrotransposons in maize, we adjusted default parameters including LTR length (100–7000 bp) and element length (1000–20000 bp). All searches required target site duplications (TSDs) of 4–6 bp (allowing one mismatch) and a 2-nt inverted motif at the terminal ends of each LTR (5' TG..CA 3', allowing one mismatch). If multiple overlapping elements were found, the one with the highest percent identity between LTRs was chosen with the '-overlaps best' option.

The resultant TE models were further annotated with LTRdigest[32], which identifies sequence features such as primer binding site, polypurine tract, and protein domains associated with previously identified retrotransposons from any organism. We used all eukaryotic tRNA entries from the UCSC gtRNA database to predict primer binding sites, and amino-acid HMM profiles of retrotransposon-associated proteins as deposited in GyDB (http://gydb.org)[34]. If RNase H, reverse transcriptase, and integrase domains were present, gene order was used to classify elements into the Ty1/Copia (integrase upstream of RNase H) and Ty3/gypsy (RNase H upstream of integrase) superfamilies.

LTR retrotransposons dominate the intergenic space of the maize genome. To capture the nested structure of these elements, generated when a newly arriving TE inserts into a TE already present at that genomic location, we computationally excised each LTR retrotransposon copy and repeated the structural search on this subtracted pseudo-genome. We repeated this computational subtraction for 80 rounds, increasing the element length by 1000 bp for each round to accommodate sequence contributed by TE fragments and TEs of other orders.

SINE and LINE: Because SINEs are transcribed by RNA polymerase III, they are often derived from one of three classes of Pol III–transcribed molecules (tRNA, 7SL, 5s rRNA). Animal SINEs of all three classes are known, whereas plant SINEs are exclusively tRNA-derived[35]. We used SINE-finder[35] to search for tRNA-derived SINEs containing RNA polymerase III A and B boxes near the polyA tail. The default A and B box consensuses (RVTGG; GTTCRA), a 25–50 bp spacer between the A and B boxes, and a spacer of 20–500 bp between the B box and polyA tail were applied. Structural SINEs were predicted only on the forward strand of the genome. LINEs were identified using TARGeT and mTEA as below for TIR elements, using LINE exemplars and 15 bp target site duplications.

TIR: Exemplar elements from the maize TE consortium (MTEC) annotation[20] were used as nucleotide queries in TARGeT[36], a pipeline designed to recover high-copy transposon and gene families. The number of elements clustered in the PHI step was increased to 10000 copies, and 200 bp of flanking sequence on either edge of genomic matches was extracted (-p_f 200). This approach recovered candidate TE sequences, but the TE boundaries and flanking sequence were unknown. To identify the boundaries of each element, we scanned each candidate and verified the presence of terminal inverted repeat (TIR) and TSD sequences indicative of the TE superfamily (see the table below), using mTEA (https://github.com/hyphaltip/mTEA/blob/master/scripts/id_TIR_in_FASTA.pl; modified to use mafft for alignment), Although TSDs and TIRs should be identical for most superfamilies upon insertion into the genome, mutations arising at the background genomic mutation rate can generate differences. Thus, we allowed mismatches of 80% of the length of a TSD or TIR to accommodate identification of these older, degraded copies.

DNA TIR TE Superfamily TSD & TIR Classification

Superfamily	TSD Length (sequence restrictions, if any)	TIR Length (sequence restrictions, if any)
DTT *Tc1/Mariner*	2 bp (TA)	13 bp
DTA *hAT*	8 bp	11 bp
DTM *Mutator*	9 bp	40 bp
DTH *Pif/Harbinger*	3 bp (TNN)	14 bp
DTC *CACTA*	3 bp	13 bp (CACTNNNNNNNNN)

In addition, MiteHunter[37] and detectMITE[38] were used to identify *de novo* structural MITEs, searching for TIR and TSDs in genomic sequences. We filtered MITE output by TSD and TIR length, and all exemplars with TIRs and TSDs of anticipated length for the superfamily were used to search using mTEA, as described above.

Helitron: HelitronScanner[39] with default parameters was deployed to identify upstream and downstream termini of helitrons, and to join upstream and downstream termini within 200–20,000 bp of each other into helitron TE copies. We predicted helitrons in both the direct and reverse complement orientations.

Family clustering: Families were identified within each superfamily of TIR TE and order of retrotransposon using the 80–80–80 rule[40], which requires that elements within a family must share 80% homology over at least 80 base pairs of 80% of the element's functional or internal domains. For LTR retrotransposons, the 5' LTR was used to cluster families, consistent with previous annotations in maize[41]. The entire element sequence was used to group TIRs, LINEs, and SINEs, because functional domains are short, and because a large proportion of non-autonomous elements lack protein-coding domains. Because the internal regions of maize helitrons are diverse and clustering methods applied to the entire element yield almost exclusively singletons[42], we used a family classification previously applied to maize helitrons that relies on 80% identity of the 30 bp at the 3' end of each copy[43], a region of hairpin-forming sequence important for rolling circle replication. All family definitions are consistent with those used previously in the maize genome sequencing project[20,41], although we implemented clustering of families in SiLiX[44]. Additionally, for each structurally defined TE in the genome, we assigned a unique identifier that indicates its superfamily and family.

Calculating genomic composition and resolving TE overlaps: As structural searches were run independently for each TE order, we filtered overlapping insertions in order to count each genomic position as derived from only one transposable element and generate a filtered set of TE annotations. As subsequent transposition into existing TEs causes them to occupy larger ranges along the genome, larger TEs are expected to be older. Since the chance of false homology increases as requirements of sequence identity are reduced, we filtered out LTR retrotransposons that occupy over 100kb along the genome, as these old large elements are more likely to be false positives. As nested insertions from most orders of TEs are known[45-48] (LTR into helitron, helitron into LTR; TIR into LTR, LTR into TIR), we retain TE copies entirely nested within another copy, but remove insertions that overlap boundaries of other copies. When copies overlap, we retain first

LTR retrotransposons, next TIR, next SINE and LINE, and finally helitrons. This removal order was chosen to favor TE orders with stronger structural signatures.

Homology Search: After a TE inserts into a position in the genome, it is subject to subsequent mutations. Because features will erode over time, making identification difficult, these changes can complicate its ascertainment by structural methods. To identify these waning TE-derived sequences, we used RepeatMasker (http://www.repeatmasker.org) to mask the B73 RefGen_v4 pseudomolecules with a repeat library consisting of structurally defined TEs. These consist of the filtered TE set described above, but with LTR retrotransposon families containing greater than 10 copies additionally downsampled to reduce computational runtime. This is necessary due to the existence of large families with tens of thousands of nearly identical copies. For these LTR retrotransposon families, we algorithmically selected exemplar elements, based on the length distribution of the TE family. Briefly, we used a Dirichlet Process Prior to identify the most likely number of normal distributions needed to generate the observed length distribution, and identified cluster membership for each element in the family. Then, we selected the copy with a length closest to the mean of each inferred normal distribution. These copies were used as exemplars in the homology search.

Comparison of transposable element annotations in v3 and v4: To compare our annotation approach with existing TE annotations generated based on homology to the MTEC repeat library (www.maizetedb.org), we annotated the AGPv3 assembly using the structural methods applied to AGPv4. We then assessed the overlap between the available RepeatMasker annotation of AGPv3 and this new annotation. This analysis revealed that only 0.6% (11,017 of 1,695,362) of LTR retrotransposons in RepeatMasker AGPv3 annotation are full-length and contain TSDs. Such striking underrepresentation is anticipated when homology-based methods are used to identify diverse TEs[49]. In addition to the improved quality of the annotation, the AGPv4 genome allows more complete reconstruction of the entire sequence of each TE. For example, we recovered 68% more Ty1/Copia and Ty3/Gypsy LTR retrotransposons with evidence of all proteins required for retrotransposition (42,929 in AGPv4 vs. 25,412 in AGPv3); in AGPv3, many of these internal domains were represented by gaps between contigs.

Diversification of maize LTR retrotransposons: To investigate the evolutionary dynamics of retrotransposition in maize since divergence from Sorghum, we applied our annotation approach for LTR retrotransposons to the *Sorghum bicolor* genome (Sorbi1). Sequences matching HMM models of RT_crm.hmm (Ty3/Gypsy) and RT_sire.hmm (Ty1/Copia) were extracted from each non-nested LTR TE they matched. As the estimated divergence time between maize and sorghum (12 Mya) predicts greater divergence than the 80% identity used to define families, generated a consensus sequence for each family using emboss cons[50] to track differences between species. We aligned these family consensuses with MAFFT mafft[51] and built a maximum likelihood phylogenetic tree with fasttree2[52]. We then collapsed sister tips on the tree if they arose from the same species, and summed the number of copies belonging to each of these species-specific lineages. Hence, monophyletic lineages of TEs, with respect to the genome they were ascertained from, are shown in Figure 2.

Data Availability: Scripts, parameters, and intermediate files of each TE superfamily are available at https://github.com/mcstitzer/agpv4_te_annotation/tree/master/ncbi_pseudomolecule

Reference

1 Berlin, K. *et al.* Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature biotechnology* **33**, 623-630, doi:10.1038/nbt.3238 (2015).

2 Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**, 563-569, doi:10.1038/nmeth.2474 (2013).

3 Nguyen, J. V. *Genomic mapping: a statistical and algorithmic analysis of the optical mapping system*. (University of Southern California, 2010).

4 Anantharaman, T. & Mishra, B. in *Algorithms in Bioinformatics, First International Workshop, WABI.* 27-40.

5 Myers, E. W. *et al.* A whole-genome assembly of Drosophila. *Science* **287**, 2196-2204 (2000).

6 Sommer, D. D., Delcher, A. L., Salzberg, S. L. & Pop, M. Minimus: a fast, lightweight genome assembler. *BMC bioinformatics* **8**, 64, doi:10.1186/1471-2105-8-64 (2007).

7 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, doi:10.1371/journal.pone.0028334 (2011).

8 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, doi:10.1534/genetics.115.177360 (2015).

9 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).

10 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

11 Haberer, G. *et al.* Structure and architecture of the maize genome. *Plant Physiol* **139**, 1612-1624, doi:10.1104/pp.105.068718 (2005).

12 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. *The Plant cell* **23**, 4280-4297, doi:10.1105/tpc.111.092163 (2011).

13 Cantarel, B. L. *et al.* MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research* **18**, 188-196, doi:10.1101/gr.6743907 (2008).

14 Monaco, M. K. *et al.* Gramene 2013: comparative plant genomics resources. *Nucleic Acids Res* **42**, D1193-1199, doi:10.1093/nar/gkt1110 (2014).

15 Wang, B. *et al.* Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nature communications* **7**, 11708, doi:10.1038/ncomms11708 (2016).

16 Soderlund, C. *et al.* Sequencing, mapping, and analysis of 27,455 maize full-length cDNAs. *PLoS Genet* **5**, e1000740, doi:10.1371/journal.pgen.1000740 (2009).

17 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, doi:10.1104/pp.114.245027 (2015).

18 Martin, J. A. *et al.* A near complete snapshot of the Zea mays seedling transcriptome revealed from ultra-deep sequencing. *Scientific reports* **4**, 4519, doi:10.1038/srep04519 (2014).

19 Keller, O., Kollmar, M., Stanke, M. & Waack, S. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27**, 757-763, doi:10.1093/bioinformatics/btr010 (2011).

20 Schnable, P. S. *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112-1115, doi:10.1126/science.1178534 (2009).

21 Vilella, A. J. *et al.* EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research* **19**, 327-335, doi:10.1101/gr.073585.107 (2009).

22 Haas, B. J., Delcher, A. L., Wortman, J. R. & Salzberg, S. L. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* **20**, 3643-3646, doi:10.1093/bioinformatics/bth397 (2004).

23 Youens-Clark, K. *et al.* Gramene database in 2010: updates and extensions. *Nucleic Acids Res* **39**, D1085-1094, doi:10.1093/nar/gkq1148 (2011).

24 Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-964 (1997).

25 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, doi:10.1104/pp.113.230144 (2014).

26 Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome biology* **5**, R12, doi:10.1186/gb-2004-5-2-r12 (2004).

27 Istrail, S. *et al.* Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc Natl Acad Sci U S A* **101**, 1916-1921, doi:10.1073/pnas.0307971100 (2004).

28 Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006-1007, doi:10.1093/bioinformatics/btt730 (2014).

29 Kumari, S. & Ware, D. Genome-wide computational prediction and analysis of core promoter elements across plant monocots and dicots. *PLoS One* **8**, e79011, doi:10.1371/journal.pone.0079011 (2013).

30 Smith, A. D., Sumazin, P., Xuan, Z. & Zhang, M. Q. DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc Natl Acad Sci U S A* **103**, 6275-6280, doi:10.1073/pnas.0508169103 (2006).

31 Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC bioinformatics* **9**, 18, doi:10.1186/1471-2105-9-18 (2008).

32 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, doi:10.1093/nar/gkp759 (2009).

33 Lerat, E., Burlet, N., Biemont, C. & Vieira, C. Comparative analysis of transposable elements in the melanogaster subgroup sequenced genomes. *Gene* **473**, 100-109, doi:10.1016/j.gene.2010.11.009 (2011).

34 Llorens, C. *et al.* The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res* **39**, D70-74, doi:10.1093/nar/gkq1061 (2011).

35 Wenke, T. *et al.* Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *The Plant cell* **23**, 3117-3128, doi:10.1105/tpc.111.088682 (2011).

36 Han, Y., Burnette, J. M., 3rd & Wessler, S. R. TARGeT: a web-based pipeline for retrieving and characterizing gene and transposable element families from genomic sequences. *Nucleic Acids Res* **37**, e78, doi:10.1093/nar/gkp295 (2009).

37 Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res* **38**, e199, doi:10.1093/nar/gkq862 (2010).

38 Ye, C., Ji, G. & Liang, C. detectMITE: A novel approach to detect miniature inverted repeat transposable elements in genomes. *Scientific reports* **6**, 19688, doi:10.1038/srep19688 (2016).

39 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 U S A* **111**, 10263-10268, doi:10.1073/pnas.1410068111 (2014).

40 Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nature reviews. Genetics* **8**, 973-982, doi:10.1038/nrg2165 (2007).

41 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).

42 Sweredoski, M., DeRose-Wilson, L. & Gaut, B. S. A comparative computational analysis of nonautonomous helitron elements between maize and rice. *BMC genomics* **9**, 467, doi:10.1186/1471-2164-9-467 (2008).

43 Yang, L. & Bennetzen, J. L. Distribution, diversity, evolution, and survival of Helitrons in the maize genome. *Proc Natl Acad Sci U S A* **106**, 19922-19927, doi:10.1073/pnas.0908008106 (2009).

44 Miele, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC bioinformatics* **12**, 116, doi:10.1186/1471-2105-12-116 (2011).

45 Gupta, S., Gallavotti, A., Stryker, G. A., Schmidt, R. J. & Lal, S. K. A novel class of Helitron-related transposable elements in maize contain portions of multiple pseudogenes. *Plant molecular biology* **57**, 115-127 (2005).

46 Morgante, M. *et al.* Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat Genet* **37**, 997-1002 (2005).

47 Jameson, N. *et al.* Helitron mediated amplification of cytochrome P450 monooxygenase gene in maize. *Plant molecular biology* **67**, 295-304 (2008).

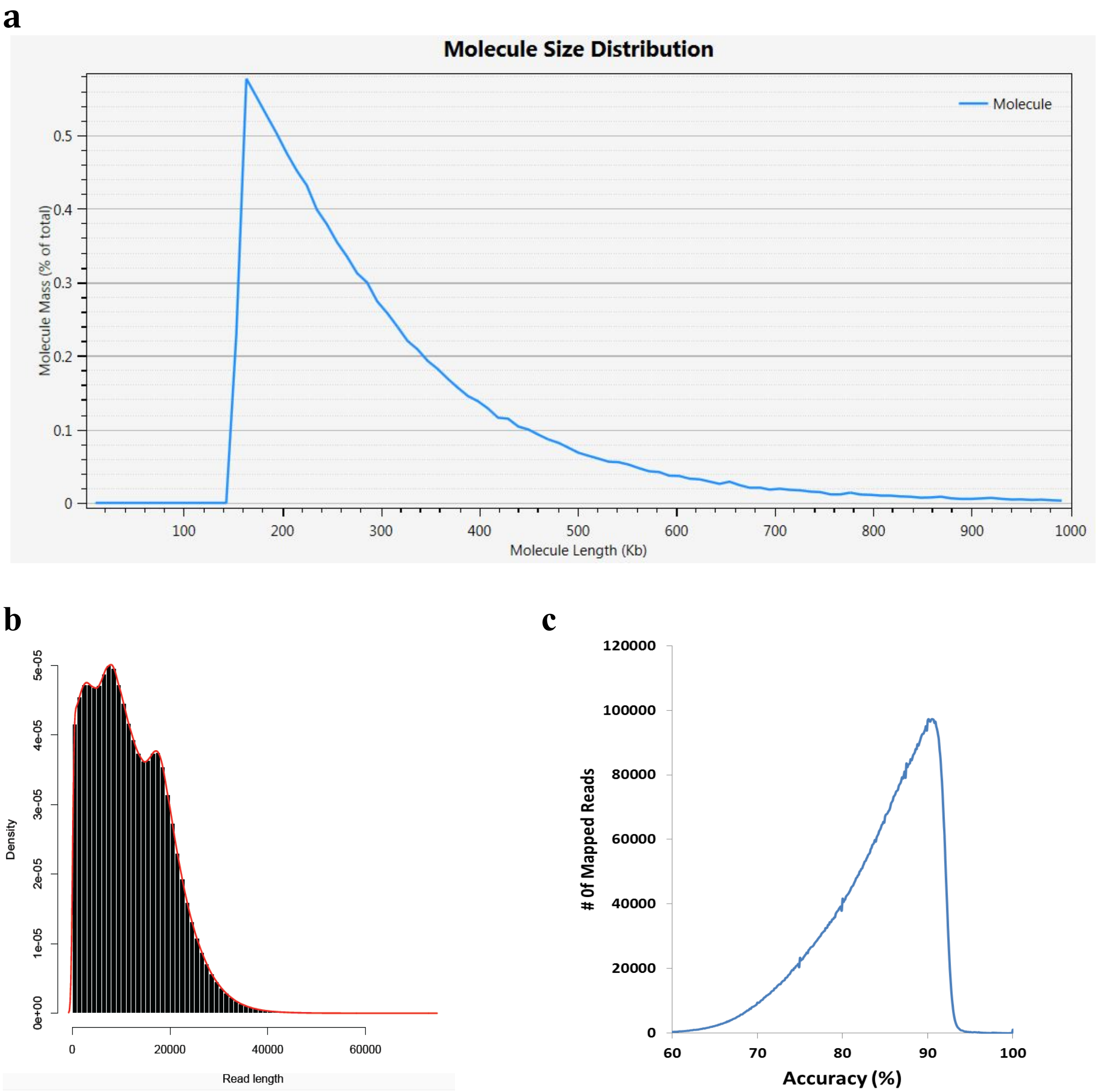
48 Jiang, N. & Wessler, S. R. Insertion preference of maize and rice miniature inverted repeat transposable elements as revealed by the analysis of nested elements. *The Plant cell* **13**, 2553-2564 (2001).

49 Platt, R. N., Blanco-Berdugo, L. & Ray, D. A. Accurate transposable element annotation is vital when analyzing new genome assemblies. *Genome Biology and Evolution*, evw009 (2016).

50 Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European molecular biology open software suite. *Trends in genetics* **16**, 276-277 (2000).

51 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**, 772-780 (2013).

52 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).



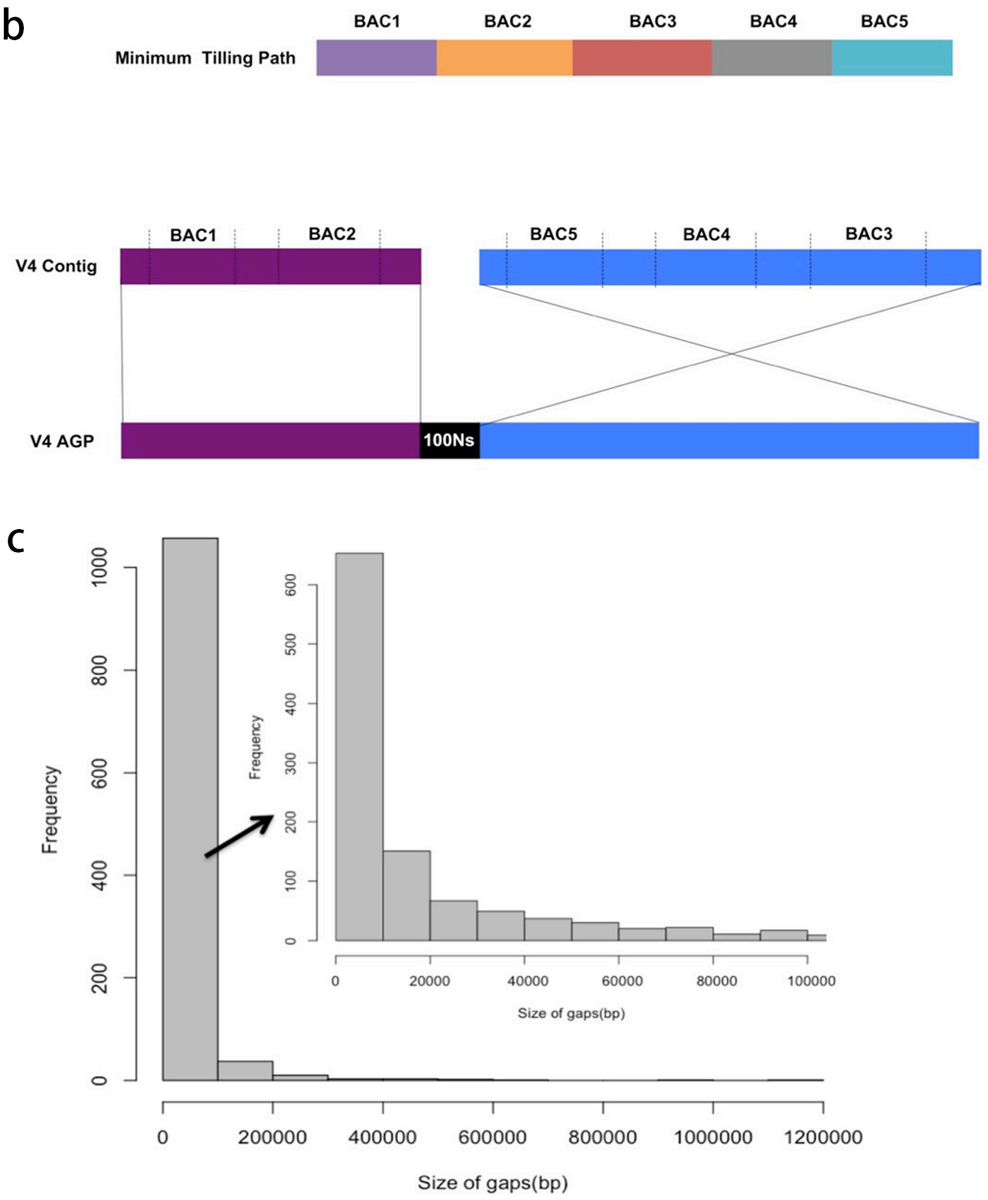
Extended Data Figure 1. Summary of data generated for genome construction. A total of 150 Gb (~60-fold coverage) of single-molecule raw data was collected for map construction. The N50 of the single molecules was ~261 kb, and the label density was 11.6 per 100 kb. After assembly, the total size of the map reached 2.12 Gb with an N50 of 2.47 Mb. 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. The sequencing error rate was estimated at 10%.

a). Size distribution of single molecules for the construction of genome maps.

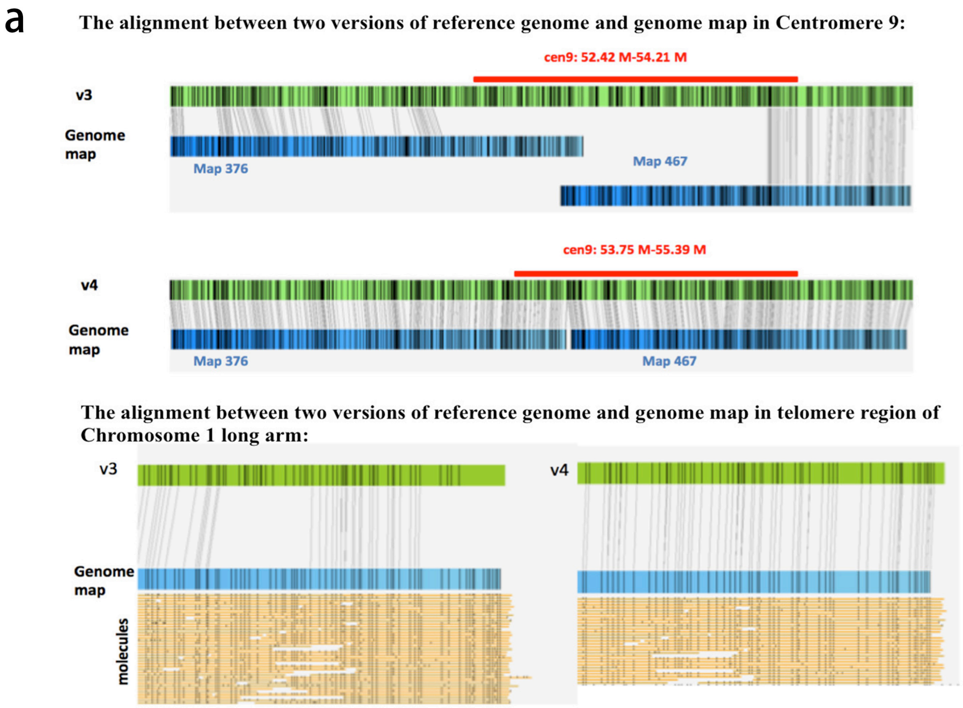
b). Length distribution of SMRT sequencing reads.

c). The accuracy of SMRT sequencing reads 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)** Contigs were ordered according to the order of the BACs the contained. **c)** Size distribution of gaps in the pseudomolecules estimated using the BioNano genome 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:

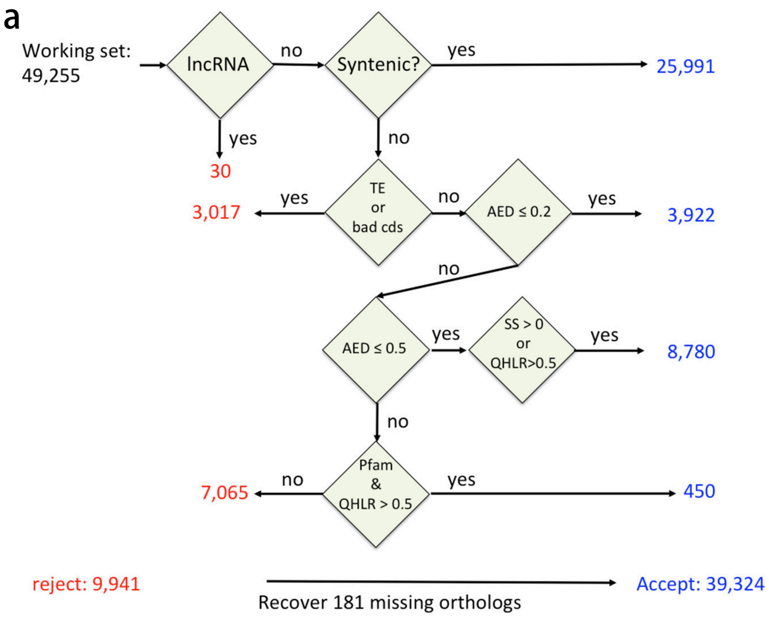
b

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
lL	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	

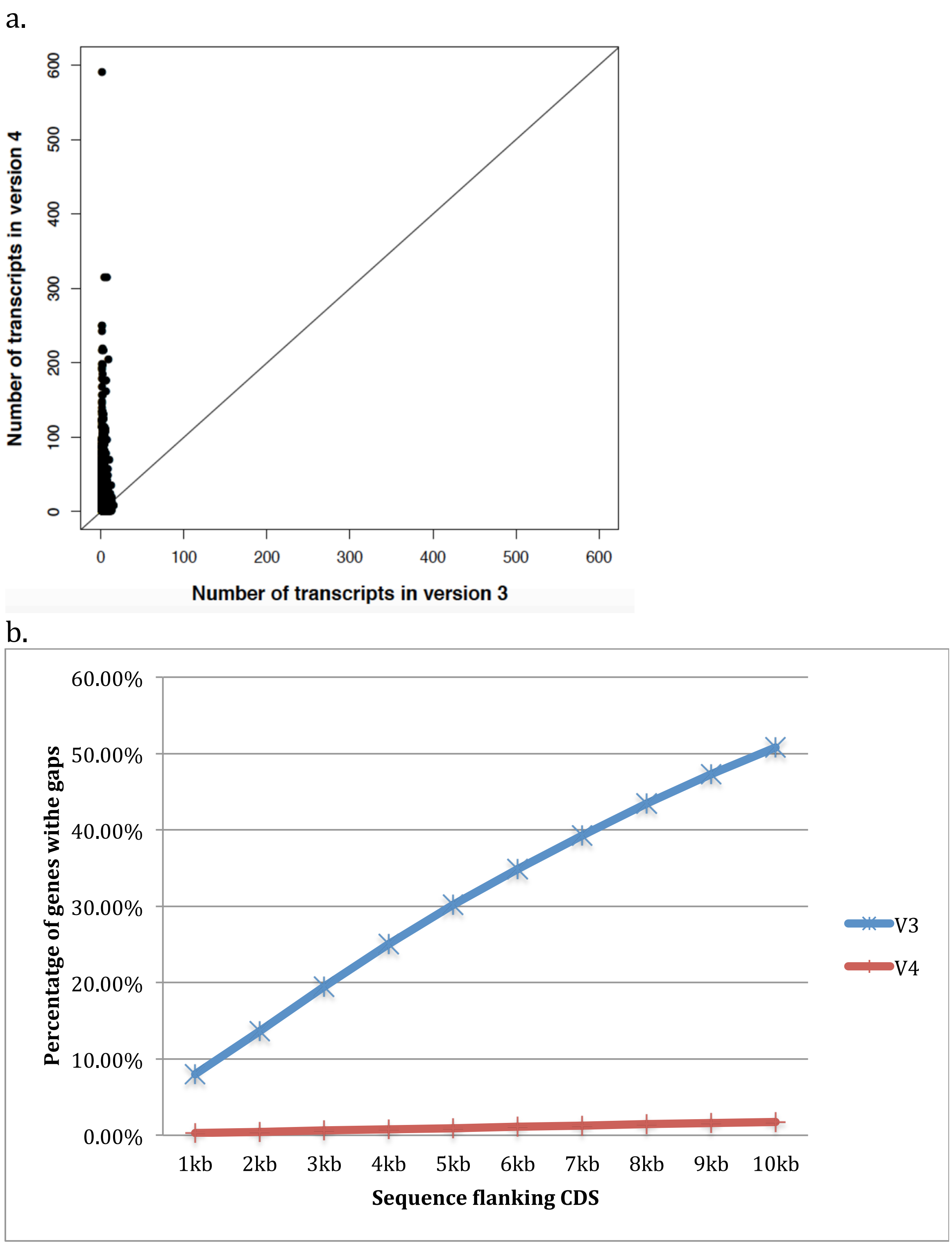
Extended Data Figure 3. Quality assessment and comparison of the assembly in centromere and telomere regions in maize B73 RefGen_v3 and v4. a) Quality assessment of centromere and telomere using genome map**. b)** Locations of centromeres on pseudomolecules defined by ChIP-seq in B73 RefGen_v3 and v4. **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 maize B73 RefGen_V4. a) The pipeline used to get high confidence gene models. **b)** Summary of B73 RefGen_v4 protein-coding gene annotation, and comparison with RefGen_v3 annotation.

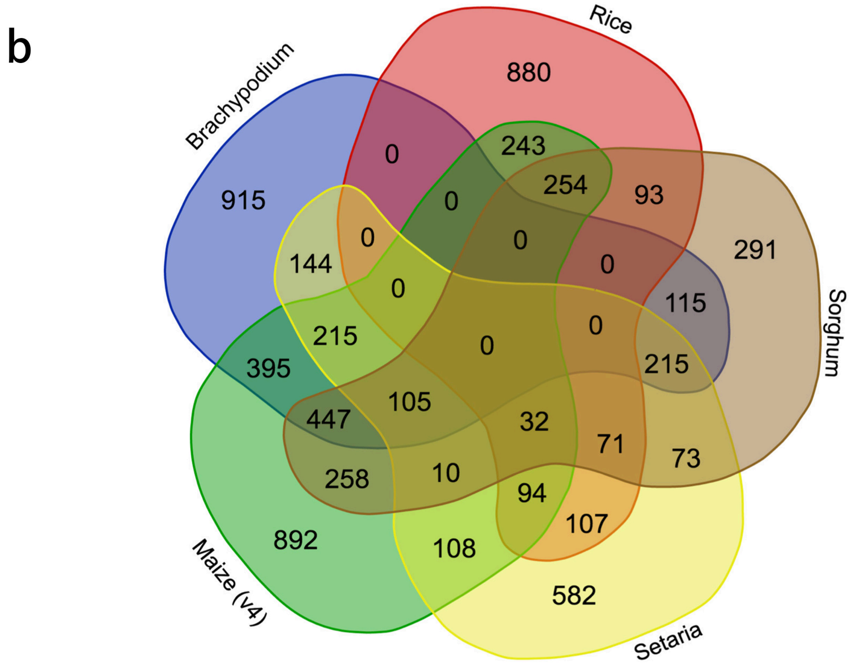


Extended Data Figure 5. Improvement of RefGen_v4 gene models. a) Number of tran of each gene in v3 and v4 annotation; **b)** Percentages of genes with gaps in flanking regio the v3 and v4 annotations.

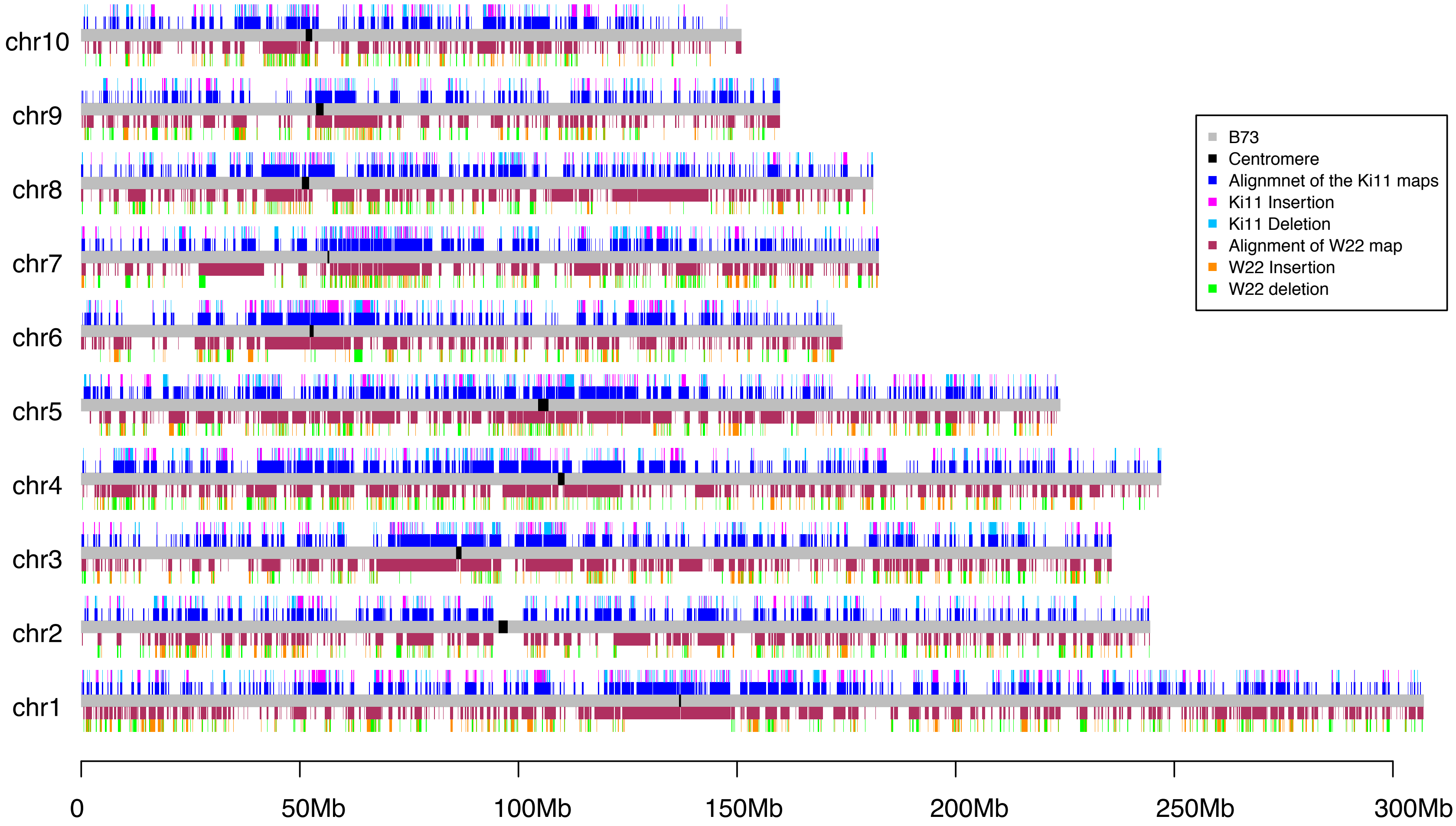
a

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)

* Number in parentheses gives the fraction (%) of ortholog sets with membership of all species and versions within the clade.



Extended Data Figure 6. Comparative analysis of the maize B73 RefGen_v4 genes with other grasses. a) Species-membership in ortholog sets, giving counts and fraction (%) of ortholog sets of which each species is a member. **b)** Venn diagram showing overlap of 6,539 ortholog sets rooted in the Poaceae (true grasses) that are deficient in gene membership among five species.



Extended Data Figure 7. Structure variation called from the Ki11 and W22 genome maps.

Extended Data Table 1. Summary of BioNano genome 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 orthologs 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 orthologs in rice, *Brachypodium*, or *Setaria* outgroup species. † Subset of background with no annotated orthologs 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%