www.sciencemag.org/cgi/content/full/326/5956/1112/DC1

Supporting Online Material for

The B73 Maize Genome: Complexity, Diversity, and Dynamics

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

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

Published 20 November 2009, *Science* **326**, 1112 (2009)
DOI: 10.1126/science.1173462

This PDF file includes

Materials and Methods
SOM Text
Figs. S1 to S18
Tables S1 to S18
References

ONLINE SUPPORTING MATERIAL

SEQUENCING, ASSEMBLY AND SEQUENCE IMPROVEMENT 2

TE SEARCH APPROACHES, DEFINITIONS OF FAMILIES, INTACT ELEMENTS AND FRAGMENTED ELEMENTS, AND GENE FRAGMENTS CAPTURED BY TES 3

CONSTRUCTION OF A LIBRARY OF REPETITIVE SEQUENCES 3
CACTA, TC1/MARINER, HAT AND PIF/HARBINGER ELEMENTS 3
Coding Elements 3
Non-Coding Elements 3
Exemplars 3
Estimation of copy numbers and genome coverage 4
Estimation of the number of intact elements carrying gene fragments 4
HELITRONS 4
LINES 4
LTR RETROTRANSPOSONS 5
MULES 5
SINES 6
GENE ANNOTATION 6
METHODS 6
Gene prediction and selection 6
RESULTS 7

ADDITIONAL ESTIMATIONS OF THE NATURE AND ERROR RATE OF GENE CONTENT 8

PREDICTIONS IN B73 REFGEN_V1 8
Estimation of the number of genes in the maize genome 8
Stringent estimation 8
Non-stringent estimation 8

FUNCTIONAL ANNOTATION OF THE FILTERED GENE SET 8

METHODS 8
RESULTS 9

ORTHOLOG AND PARALOG DETERMINATION 9

NOTE ON LINEAGE-SPECIFIC GENE FAMILIES 9

RNA-SEQ 10

Leaf transcriptomics 10
Transcriptomics of shoot apical meristems (SAMs) and seedling from reciprocal hybrids 10
RESULTS 10

CENTROMERE METHODS 11

IDENTIFICATION AND DRAFT SEQUENCING OF CENTROMERIC BACS 11
ENRICHMENT AND REPRESENTATION CALCULATIONS 11
MAPPING CHIP SEQUENCES TO BACS OR B73 REFGEN_V1 12
IDENTIFICATION AND MAPPING OF CENTROMERIC EST AND CDNA SEQUENCES 12

SYNTENY ANALYSIS AND IDENTIFICATION OF THE MAIZE LINEAGE WHOLE GENOME DUPLICATION 12

METHODS 12
RESULTS 13
PREFERENTIAL RETENTION OF GENE FUNCTIONAL CLASSES 13
PREFERENTIAL GENE LOSS IN ANCESTRAL HOMOEOLOGOUS CHROMOSOMES 14
IDENTIFICATION OF PARALOG CLUSTERS 14

MAIZE FINISHING GUIDELINES **14**
RULES FOR DETERMINING REGIONS TO BE FINISHED IN THE MAIZE GENOME: 14
GUIDELINES FOR IMPROVEMENT OF THE MAIZE GENOME: 15
SUPPORTING FIGURES **18**
SUPPORTING TABLES **20**
SUPPLEMENTARY REFERENCES **22**

Sequencing, Assembly and Sequence Improvement

The maize genome was sequenced via a BAC-by-BAC approach with a minimum tiling path (MTP) of 16,848 BACs derived from an integrated genetic and physical map (*S1*). Clones were picked, as described (*S2*). Sheared DNA from each clone was ligated into the pSMART (Lucigen, Middleton, WI) plasmid vector. Each BAC library received 2X384-well paired end sequences, resulting in ~4- 6X coverage. Data were assembled, confirmed by BAC End Sequence, checked for minimum coverage standards (submitted to GenBank as HTGS_FULLTOP), and sent for automated sequence improvement. Prior to sequence improvement, fosmid end sequences, from a repository composed of fosmids prepared and sequenced at the Washington University Genome Center, were added to the assemblies. The fosmid clones were chosen by running a script against the shotgun assembly that uses BLAST to compare 600 bp segments of the assembly against the maize sequence read repository database. If at least 99% identity was noted, the program retrieved the fosmid (and fosmid mate pairs, if available) and incorporated it into the assembly to enhance order and orientation. Consensus sequences were evaluated by a K-mer analysis to identify repetitive regions (*S3*). Automated improvement involved directed sequencing across gaps and low quality sequences within non-repetitive regions only (submitted to GenBank as HTGS_PREFIN). Predicated on the average number of bases finished per submitted clone (26,968 bps), as of February 1, 2009, a total of 379,598,676 base pairs of finished unique sequence were submitted to GenBank.

Following automated sequence improvement, additional data downloaded from GenBank, such as cDNA sequences and sequences from subtractive libraries with methyl-filtered DNA and high C_0t techniques, were incorporated into the assemblies (submitted as HTGS_ACTIVEFIN). Manual improvement was performed on non-repetitive regions only, with guidelines established by the MGSC (see Supplemental Note: Maize Finishing Guidelines). Improved sequences were submitted to GenBank as phase-I improved (HTGS_IMPROVED).

The B73 RefGen_v1 also includes published sequences from 55 B73 BACs that were not generated by the maize genome sequencing project: Genbank Accession numbers: AC147602, AC147791, AC148152, AC148167, AC148479, AC149475, AC149478, AC149633, AC149640, AC149810, AC149818, AC149828, AC149829, AC150739, AC152495, AC155352, AC155363, AC155376, AC155377, AC155383, AC155397, AC155417, AC155434, AC155496, AC155507, AC155537, AC155610, AC155622, AC155624, AC159612, AC166636, AF466202, AF546187, AF546188, AY211534, AY211535, AY530952, AY542798, EF517600, EF517600, EF562447 (*S4*-*S12*). Note that some accession numbers reference more than one BAC.

In total, the B73 RefGen_v1 contains 2,048 Mb in 125,325 sequence contigs (N50 of 40 kb), forming 61,161 scaffolds (N50 of 76 kb) of the maize genome, which consists of an estimated 2.3 Gb (*S13*). We thus estimate that ~250 Mb (~10.8%) of the genome is missing from B73 RefGen_v1. The ~7% of the genome (~170 Mb) that is not contained within the maize physical map accounts for ~70% of the missing sequence. Some of the remaining missing sequence can be attributed to tandem repeats as illustrated in Table S13, which shows that 90% of the estimated 30 Mb of knob 180 repeat are missing in the assembly, as well as 80% of the 2 Mb knob 350 repeat, 45% of 3 Mb CentC and 86% of the estimated 35Mb of 45S rDNA, among others. Thus these analyzed repeats alone account for a total of 60 Mb of missing DNA (27+ 1.6 + 1.35 + 30 Mb). Because each

BAC was sequenced at 4-6x coverage, it is possible that some sequences were missed in DNA sequencing. We also cannot exclude the missing DNA from an assembly-based collapse of two highly similar LTRs of a recently inserted retrotransposon. Comparison of B73 RefGen_v1 to ten highly curated gene sequences showed that the maximum sequencing error rate was 0.025%, as summarized in Table S1.

TE Search Approaches, Definitions of Families, Intact Elements and Fragmented Elements, and Gene Fragments Captured by TEs

The 2045 Mb of maize (B73) genomic sequence (excluding gaps) analyzed in this study (B73 RefGen_v1) were downloaded from The Maize Genome Sequence Browser (http://maizesequence.org) and TEs were annotated with a variety of approaches.

Construction of a library of repetitive sequences

To systematically identify repetitive sequences, the maize genomic sequence was downloaded and clustered with RECON, a program for *de novo* identification of repetitive sequences (*S3*). The cutoff for consideration as a repetitive sequence was 10 or more copies in the final sequence set. The resulting library (containing 33,201 repetitive sequences) is referred to as the RECON library below. It is used for the identification of *Mutator*-like elements (MULEs) and for other DNA elements that were absent from other collections (see below).

CACTA, Tc1/Mariner, hAT and PIF/Harbinger elements

Coding Elements

Coding regions of the DNA TE superfamilies *PIF/Harbinger*, CACTA, *Tc1/Mariner* and *hAT* were identified with consensus sequences derived from the most conserved (catalytic) region of each superfamily, and TBLASN searches were performed with a pipeline called TARGeT (*S14*). Retrieved sequences were aligned with MUSCLE (*S15*) and phylogenetic trees generated for each superfamily with MEGA4 and PAUP version 4.0b8 (*S16, S17*). In addition, the conserved sequences of previously identified maize DNA TEs (*Ac*, *En*, *Doppia*4, *PIF* and *Bergamo*) were also included as queries.

Next, full-length coding elements [including terminal inverted repeats (TIRs) and target site duplications (TSDs)] were determined with a newly developed TIR/TSD structure-based method that includes a statistical method to filter false positives. Full-length copies were further confirmed with BLAST to identify homologs in the genome.

Non-Coding Elements

Identification of full-length coding elements is a prerequisite for detecting non-coding elements that are deletion derivatives. Therefore, full-length coding copies were used to survey the maize genomic sequences for related non-coding elements including MITEs via BLAST and RepeatMasker (http://www.repeatmasker.org/). As with the coding elements, non-coding elements previously identified in maize [including *Dotted* (*rDt*), *Ds*, *Ds1*, and the MITEs *Hbr* (*Heartbreaker*), *Irma/mPIF*, *TouristZm1*, *StowawayZm1]* were also included in this screen.

Exemplars

To reduce the redundancy of recovered sequences and to hasten future annotation of maize genomic DNA, we generated a collection of exemplars (representative TE sequences) using the following procedure. All element sequences from the same superfamily were compared with BLASTN. The element with the most matches (cutoff at 90% identity in 90% of the element length) was considered as the first exemplar. Thereafter, this element and its matches were excluded from the group and a second round BLASTN search was conducted with the remainder of the elements, leading to the generation of the second exemplar. This process was repeated until all elements were excluded. For coding elements in these four superfamilies, a phylogenetic tree was generated for each family. On the basis of visual examination of the phylogenetic tree, a full-length element was chosen from each clade as the exemplar. The exemplars for both coding and non-coding elements then were used to mask the RECON library (with RepeatMasker) and the unmasked sequences were examined for elements with features of CACTA, *hAT*, *Tc1/Stowaway*, and *PIF*/*Tourist* superfamilies. This led to the identification of additional elements that were not included by the exemplars. Each exemplar and each additional element identified from the RECON library were considered a family.

Estimation of copy numbers and genome coverage

The exemplars and sequences derived from the RECON library were used to mask B73 RefGen_v1 and the output of RepeatMasker was used for estimation of copy number and genomic coverage of each superfamily. The redundant matches in the output were eliminated by excluding the shorter match (for copy number calculation) if two elements matched the same region and the overlapped part was 90% or longer of the shorter match. If an element in the genomic sequence matched an exemplar over the entire sequence, or if the truncation was less than 20 bp on each end, this element was considered to be an intact element. Otherwise it was considered as a truncated element or half of a copy. Fragmented elements that lack both ends (truncated more than 20 bp on both ends) were not included in copy number estimation. The genome coverage of TEs was estimated as the total sequence masked by each superfamily.

Estimation of the number of intact elements carrying gene fragments

To identify elements with gene sequences in their internal regions, the sequences of exemplars were used to mask B73 RefGen_v1. Candidate elements were retrieved if they possess terminal sequences of exemplars from the same super-family at both ends, less than 15 kb, contain non-exemplar sequences, and with a minimum of two copies in the genome. The sequences of candidate elements were used to search against the coding region of the filtered gene set of B73 RefGen_v1, and the genes that align with the elements with at least 100 bp in length were considered candidate parental genes. To minimize the effect of gene annotation artifacts, captured gene fragments were excluded if they match examplar sequences (>= 30 bp) of any type of TEs at the nucleotide level, or match any known transposases (> 50 amino acids) at the protein level, or are flanked by TE terminal sequences. The remaining gene sequences were used to search against plant proteins in NCBI and a maize EST database, downloaded on Sep. 6, 2009. Only those genes matching known plant proteins ($E < 10^{-5}$) and matching a maize EST sequence (at least 100 bp) are considered as final parental genes. The elements with one or more corresponding final parental genes are considered as TEs carrying gene fragments.

Helitrons

Helitrons were sought in the B73 RefGen_v1 by searching for the canonical 3' and 5' ends associated with intact elements, and requiring at least two independent elements with these exact ends to confirm that they are members of a unique *Helitron* family, as described (*S18*). Because many *Helitron* families in some plant genomes are predicted to be present with zero or one intact members (*S18*), this approach provides a minimum estimate of total *Helitron* copy number. A family is defined as all *Helitrons* with the same 3' end (>80% identity over the terminal 30 bp). Intact members of the same family have both 3' and 5' ends, while fragmented elements are defined as all non-intact elements with at least 100 contiguous bp of >80% homology to an intact *Helitron* (*S19*). A minimum total number of *Helitrons* (fragmented plus intact) in B73 RefGen_v1 was calculated by counting the number of 3' ends (~21,000) and 5' ends (~22,000) for the eight identified families. The ratio of apparently fragmented to intact elements in the B73 RefGen_v1 is greater than ten to one, but this fragment excess is at least partly an annotation artifact caused by the great number of gaps, incorrectly ordered scaffolds and improper assemblies in the repeat-rich regions of the sequenced BACs of of B73 RefGen_v1 (*S19*).

LINEs

Candidate Long Interspersed Nuclear Elements (LINEs) were identified in the maize genome sequence with a Perl script that searched for host site duplications of a given size range flanking a block of sequence of appropriate length that terminated on one end with a simple sequence repeat, usually poly A (this Perl script is available upon request from Phillip SanMiguel, Purdue University). This large pool of sequences was filtered by requiring that candidates encode residues of reverse transcriptase generally shared among LINEs, but not among LTR retrotransposons. A more detailed description of this strategy is provided in (*S20*). Candidates were grouped into families of elements sharing >80% identity over >80% of their length, as detailed in (*S21*). Exemplar sequences for each family were manually chosen to encompass the majority of the sequence variation of the family and possess the most intact long open reading frames. The percent of the B73 RefGen_v1 composed of LINEs as well of counts of LINEs, derives from a RepeatMasker run with the exemplars used as a custom library.

LTR retrotransposons

Full length LTR retrotransposons (those with both LTRs) were initially identified by structural criteria, with several iterations of masking and reinvestigation, as described (*S20*). Families were defined with the criterion of >80% identity over >80% of their length (*S21*) specifically in analysis of the 5' LTR of intact elements. Exemplars were also identified by LTR homology and novelty criteria, and these were used in subsequent homology searches. As with *Helitrons* and any other large TEs, the B73 RefGen_v1 provides an aberrantly high value for fragmented elements because of the fragmented nature of the draft sequence (*S20*). All matches from the RepeatMasker output (with default parameters) were considered elements. The fragmented state of B73 RefGen_v1 decreases the numbers of detected intact elements relative to their true numbers, but it can lead to an inability to detect many LTR retrotransposons that have only a few (e.g., one) intact family members.

MULEs

MULEs are *Mutator*-like elements that have acquired internal sequences derived from genes. To identify members of the *Mutator* superfamily, repetitive families in the RECON library were first searched against known TEs from RepBase (http://www.girinst.org/). Families that showed significant similarity at the nucleotide or protein level (BLASTX or BLASTN, E< 10^{-10}) to known MULEs were considered to be a MULE. If a repetitive sequence was not similar to any known TE, it was defined with the following procedure. The relevant sequence was used to search the maize genome database and at least 10 matches (BLASTN, E< 10^{-10}) and their 100 bp (or longer if needed) flanking sequences on each side of the matches were recovered. The recovered sequences were then aligned with "dialign 2" (*S22*), and the resulting output examined for the presence of possible borders between putative elements and their flanking sequences. A border was defined if the sequence homology stops at the same position for more than half of the aligned sequences, and the sequences at the termini of the putative elements were compared with known TEs (*S21*). Furthermore, the sequence immediately flanking the "border" was examined for the presence of TSDs. For example, a MULE usually has a 9-bp TSD flanking the element. Some of the repetitive sequences in the RECON library represent full-length elements, which were directly used for analysis. Other sequences in the RECON library are fragmented elements. In this case, a full-length element, which is most similar to this sequence was used for subsequent analysis. Pack-MULEs, *Mutator*-like elements carrying gene fragments, were identified similarly as for CACTA, *Tc1/Mariner*, *hAT* and *PIF/Harbinger* elements that carry gene fragments (see above) with the following modifications: 1) the element should be less than 20 kb; 2) the matching for EST sequence was not required for parental genes; 3) the requirement for two copies was not applied to candidate elements. Instead, the presence of TSD (8-12 bp) was verified for individual candidates to ensure that they are legitimate elements.

For some MULEs, both internal and terminal sequences are conserved. If such elements shared at least 80% identity in 80% of the element length, they were considered as a family. For elements with highly variable internal regions, such as Pack-MULEs, a family referred to elements with at least 80% similarity in their terminal region. The estimation of copy number and genome coverage for MULEs was conducted similarly as for CACTA, *Tc1/Mariner*, *hAT* and *PIF/Harbinger* elements (see above).

SINEs

Short Interspersed Nuclear Elements (SINEs) were identified by their structural properties and their repetition, with families defined by different sequences and origins, as described (*S20*). Because of their small size, only intact elements were counted.

Table S2 summarizes the results of TE discovery described in main text.

Gene Annotation

Methods

Gene prediction and selection

Genes were predicted on the basis of assembled contigs from 16,006 BAC clones with a combination of the Gramene evidence-based gene build pipeline (*S23*) and FGENESH (*S24*). For a small subset (506 BACs) only FGENESH was used. Prior to annotation, sequences were masked with MIPs REdat v4.3 library (*S25*) which resulted in masking of 78.4% of genome sequence. The gene-build incorporated sequence evidence from both maize and other plant sources available as of October 2008 as follows:

a) Maize full-length cDNAs (FLcDNA): 14,097 from the Arizona Maize Full-length cDNA Project (*S26*); 36,430 from Ceres (*S27*)
b) EST: 2,000,333 maize; 1,217,859 rice; 2,448,641 other monocots
c) mRNA: 18,181 maize; 72,919 rice; 14,015 other monocots
d) Proteins:
  a. 359,942 from Swiss-Prot from all species
  b. 494,444 non-maize plant from Trembl
  c. 94,734 GenBank proteins from plant species
  d. 52,177 rice proteins from rice gene annotations (*S28*)
  e. 36338 proteins from sorghum gene annotations (*S29*)

For many genes, multiple spliced transcripts were preserved with high confidence cDNA/EST support (at least 99% sequence alignment identity). The resulting gene set was filtered by translation length: 50 amino acid residues for cDNA or multiple-EST supported genes, 25 residues for protein-supported genes and 100 residues for single-EST supported genes. FGENESH models were incorporated into evidence-based predictions when the former could extend the open reading frame of an otherwise incomplete coding sequence. FGENESH models that did not overlap an evidence-based prediction were used "as-is." The resulting BAC-level annotations were projected onto the reference chromosome sequence on the basis of coordinates in the Accessioned Golden Path (AGP). This removed redundant annotations due to overlap between adjacent clones in the tiling path. Some genes failed to project because their models were disrupted by assembly breakpoints. These were re-annotated directly on the chromosome assembly.

The resulting initial "working set" included 109,563 annotated loci. These were subjected to the protein annotation pipeline (see below) and to the EnsemblCompara GeneTree pipeline (*S30*) and resulting annotations, as well as additional screens, were used to remove likely false positives and transposon-encoded genes. Genes were required to encode a product of at least 50 aa and either have FLcDNA evidence or show evidence of homology (BLASTP E-value ≤ 1e-10) with annotations from Arabidopsis (TAIR8)(*S31*), sorghum (Sbi1.4 "high" confidence set)(*S29*), or rice (TIGR Release 5, non-transposon)(*S28*). Despite the fact that DNA was substantially masked prior to gene-calling, many working set genes were detected as transposable elements after screening against novel low-copy elements from both the Class I & II, as well as *Helitrons* and Pack-MULES. Specifically, if a gene was found to be within a Pack-MULE or a *Helitron*, or >70% of the coding region matched a TE exemplar or other TE sequences (see "TE Search Approaches and Definitions of Families, Intact Elements and Fragmented Elements"), it was excluded. Additional transposons were identified on the basis of InterPro domains (see Table S3) as well as manual screening of large maize-specific families. Genes lacking complete CDS or having a length less than half that of its best significant hit (in either of rice, Arabidopsis or sorghum) were regarded as likely pseudogenes with the following exception: those showing synteny (see below) and lacking a corresponding homoeologous copy were included as possible real, but missanotated, genes. This affected 1,308 annotations in the filtered set.

Results

The maize genes were predicted by combination of evidence-based predictions with *ab initio* Fgenesh predictions with the Gramene gene build pipeline (*S23*). The evidence used to predict genes includes species-specific or cross-species full-length cDNAs, ESTs, and proteins. The gene prediction was first conducted on individual contigs in the sequenced BACs. The raw gene set was projected to AGP and filtered on the basis of TE, orthology to rice/sorghum, protein length and CDS completeness, and evidence type (see Fig. S2). Finally, 32,540 genes were selected for gene related analysis. Not all genes predicted on the basis of the BACs have been successfully projected to AGP since some of these genes cross a contig break point in the assembly. On the other hand, due to incompleteness of the evidence used for this version of the gene build (eg, >10,000 FLcDNAs from Arizona were not included), and the improved genome sequence quality (AGP vs. contigs), these 32,540 genes represents an incomplete gene set on AGP. Previously, it has been estimated that maize genome contains 42,000-56,000 genes (*S5*) or ~50,000 transcripts (*S27*). Our analysis confirmed that the gene number is likely at or below the lower boundary of the first estimation. The AGP repeat-masking ratio is ~79.7% with either MIPS or the Maize TE Consortium (*S32*) repeat library. With an AGP size of 2095Mb, the non-repeat region is about 425Mb, which is larger than that in both the rice (252Mb) and sorghum (309Mb) genomes (*S29*). With an average gene length 3733bp, the total genic region is ~121Mb (28.6%), and the average gene density was 76.5 genes/Mb in the non-repeat regions. Compared to genes in rice/sorghum/Arabidopsis, the maize genes have similar exon size, larger intron size, and higher GC content (Table S5, Figs. S3-S7). Larger intron sizes can be attributed to insertions of transposable elements (Table S6)(*S5*).

It is interesting to note that the average protein length of the maize genes is smaller than their orthologs in rice/sorghum/Arabidopsis (Table S5; Fig. S8), possibly due to the incompleteness of the genome at the intra-BAC contig level (average size ~15kb, median size ~7.4kb). As a comparison, an independent gene-build was performed solely on the basis of a collection of 63,851 FLcDNAs (*S27, S28*), which generated 20,867 genes having coverage of at least 95% and alignment identity of at least 99% with FLcDNA. As shown in Table S7, these annotations have shorter translations on average than the whole filtered set. Several factors may have contributed to the short length of maize CDS: 1) incompleteness of the genome sequences at the intra-BAC contig level, which leads to partial or split genes, 2) incompleteness of the FLcDNAs (see Table S7). (*S3*) *ab initio* predictions, which are used heavily in rice/sorghum gene annotations, tend to have more exons than evidence-based genes. As a comparison of FLcDNAs, rice FLcDNA length: average=1746bp, median=1,543bp; maize FLcDNA length: average=1,220bp, median=1,187bp; from Arizona: average=1,441bp, median=1,420bp.

To estimate coverage of the gene space, 63,851 full-length cDNAs (*S27*) were aligned to B73 RefGen_v1 on the basis of >95% sequence identity. Approximately 95% of these cDNAs were successfully mapped. Presuming cDNA authenticity and that all genes have equivalent distributions in sequenced and non-sequenced regions, and considering that some cDNAs mapped only partially to the genome, it appears that 5-9% of B73 genes are not included in B73 RefGen_v1. As an example, 16 well-studied starch biosynthesis genes were mapped in detail, as summarized in Table S8. Of these, 12 aligned perfectly or nearly perfectly with B73 RefGen_v1. One gene mapped to two different contigs, thus generating two partial genes, but aligned perfectly to B73 RefGen_v1. The sequence assembly, therefore, improves the quality of gene prediction. Three genes from the test set do not align in 23-50% of their nucleotides, suggesting a relatively small degree of gene sequence incompleteness. The specific test set comprises large cDNAs (~2.6 - 3.3 kbp, compared to ~1.2 kbp average), so they may be more likely on average to truncate when mapped to intra-BAC contigs.

Additional Estimations of the Nature and Error Rate of Gene Content

Predictions in B73 RefGen_v1

In initial releases of fully automated angiosperm genome sequences, the accuracy of gene prediction has typically been incorrect by anywhere from 15% to more than 50% (*S33*). This has been primarily caused by the great complexity of angiosperm genomes, especially the abundance and diversity of TEs and gene fragments (*S33, S34*), which leads to inflated gene number predictions, especially lineage-specific genes. Given the fact that the maize genome has an order of magnitude more Mb of TEs than any previously sequenced plant genome, and a great wealth of gene fragments (*S34*), it was expected that the dependability of the maize gene annotations would be exceedingly low. However, because the TE annotation presented in this manuscript is comprehensive, it was hoped that this intrinsic inaccuracy could be minimized. With the manual annotation criteria previously described as a model for determining the accuracy of any sequenced plant genome (*S33, S34*), we investigated 200 randomly chosen filtered gene models from B73 RefGen_v1. Remarkably few of these genes (only two) were found to have characteristics suggesting that they were encoded by unrecognized TEs. Comparison of this 1% value, for instance, to the almost 50% of initially predicted genes that turned out to be TEs in the rice genome (*S35*), indicates how effective and comprehensive the TE discovery process has been for B73 RefGen_v1.

Additional studies that compare annotations within the filtered gene set to established gene structures are given in Tables S8 and S9.

Estimation of the number of genes in the maize genome
The collection of 63,851 FLcDNAs (*S27, S28*) were mapped to the AGP with BLAT (≥95% alignment identity), The mapped FLcDNAs with coverage>30%, identity≥95% is ~94.7% (this is equivalent to the gene space coverage of the maize sequence). ~90% of mapped FLcDNAs have coverage >90%. Therefore, if we assume all genes have an equivalent distribution in the sequenced and unsequenced regions, the genic region coverage of the current AGP is >0.947*0.9+0.947*0.1*(0.9+0.3)/2=91.3%. This is most likely a low estimate since we assume all FLcDNAs are real.

Stringent estimation
Among the FLcDNA-based genes, 18,329 genes were covered by the Filtered Set. Therefore, if we assume all these FLcDNA-based genes and the Filtered Set are real, then the total gene number in ZmB73v1 is 32,540 * 20,867 / 18,329 = 37,045

Non-stringent estimation
In (*S27*), the transcript number is estimated to be ~50,000. Due to the high redundancy in the transcripts used in the study, the Ceres cDNA set (~36,000) is mapped to <24,000 gene loci. Thus, the maize gene number is estimated to be ~33,333. In the Filtered Set, 85% of genes are supported by cDNAs/ESTs. Assuming that the clonable genes in the cDNA libraries comprise 85% of all genes in maize, the real maize gene number on ZmB73v1 on the basis of these criteria is 39,211.

Our analysis confirms that the final gene count is likely to be at or below the lower boundary of the first estimation of 42,000 genes (*S5*). New gene predictions will be performed on the next version of the assembly by incorporating all up-to-date evidence support; this will give a more accurate estimation of the gene count in the maize genome.

Functional Annotation of the Filtered Gene Set

Methods

In order to assign putative function to the 32,540 genes in the Filtered Set, the gene predictions were analyzed through a computational pipeline that extends the Ensembl protein annotation pipeline (*S36*). In the first phase, translations were annotated with InterPro protein domains (*S37*) with standard modules (Ncoils, Pfam, PIRSF, Prints, Profile, Prosite, Seg, Superfamily) in addition to modules such as TMHMM (*S38*) and SignalP (*S39*). In the second phase, known as cross-reference (Xref) mapping, gene predictions were associated with canonical proteins and mRNAs from established databases, on the basis of best-in-genome alignment between transcribed or translated sequence predictions as appropriate with Exonerate (*S40*). Databases whose maize entries were directly aligned include UniProt (SP-TREMBL, SWISSPROT, VARSPLIC), RefSeq, UniGene, as well as gene loci with known sequence compiled by MaizeGDB. A further step projected GO terms and EntrezGene identifiers onto the gene predictions via corresponding InterPro or Xref assignments.

Results

Of 32,540 predictions in the Filtered Gene Set, 30,626 (94.1%) were annotated with at least one protein domain and 22,847 (70.2%) were assigned at least one InterPro ID. Table S11 lists the 50 most common InterPro domains that were assigned within the Filtered Gene Set. 21,771 (66.9%) predictions matched at least one canonical protein via name-based lookup and an overlapping set of 17,430 (53.6%) predictions matched at least one canonical protein on the basis of the sequence alignment, yielding a total of 25,765 (79.2%) filtered genes that are associated with an independently annotated protein. 66,719 Gene Ontology (GO) terms were annotated in the Filtered Gene Set. The peptides of 17,301 (53.2%) canonical prediction transcripts were assigned at least one GO term, with an average of 3.9 terms per prediction (Fig. S9). The 10 most common GO terms in the Filtered Gene Set contains for each functional class, describes the distribution of three classes of gene function in the Gene Ontology and lists the top 10 terms for each class. Of 4,151 known gene loci that were retrieved from GenBank, 2,280 (54.9%) were successfully mapped to the Filtered Gene Set by the Ensembl Xref pipeline.

Ortholog and paralog determination

Orthologous and paralogous relationships between protein-coding genes of maize, sorghum, rice, and Arabidopsis were assigned with the EnsemblCompara GeneTree method, as described (*S30*). In this approach evolutionary histories (duplication and speciation events) are inferred by reconciling gene trees constructed for each protein family with the established species tree topology. Input for this analysis was the longest translation for each gene locus, filtered for transposons and other low-confidence genes, from the following genome annotation resources: the maize filtered set, the sorghum genome (*S29*) JGI release Sbi 1.4 from March 2008, The Arabidopsis Information Resource (*S31*) release 8 from April 2008, and the MSU/TIGR Rice Genome Annotation Resource (*S28*) release 5 from January 2007. Clustering was performed by an all-vs-all BLASTP followed by the extraction of genes linked either by best reciprocal BLAST, or BLAST score ratio above a threshold of 0.33. The resulting 15,741 clusters were subjected to multiple sequence alignment, tree-building, and ortholog/paralog calling as detailed in (*S30*).

Note on Lineage-Specific Gene Families

Comparisons of gene families among grasses and Arabidopsis revealed inconsistencies in annotation quality that manifested as lineage-specific and species-specific families. The excess of rice-specific families can be explained by an excess of likely false-positive gene calls in the TIGR Release 5 set (*S28*). This is indicated by numbers of families comprised of genes annotated as “hypothetical” (lacking extrinsic evidence). Overall, there were 1,664 such families, representing ~13% of the 13,055 rice families. However 758 of these families were rice-specific, representing 68% of the 1,110 rice-specific families. We also examined more closely the rice/sorghum-specific families, which surprisingly exceeded the number of maize/sorghum-specific families. Here, we found a combination of explanations: a) TE families that were successfully filtered from the maize annotation but were still present in the rice and sorghum gene sets; b) genes absent from maize due to incomplete sequence; and c) genes absent from maize due to incomplete annotation (See Table S12).

RNA-Seq

Methods

Leaf transcriptomics

Plants were grown under a 80:20 mix of metal halide: capsylite halogen lamps at a fluence of 500 umol/m2/sec, 12L:12D, 31ºL, 22ºD, 50% rel. humidity. Total RNA from 1 cm segments of leaf tip and base tissue from the third leaf of 9-day-old (DAP) Mo17 seedlings was extracted with Trizol reagent (Invitrogen, Carlsbad, CA). Approximately 100 ng mRNA was isolated with the Dynabeads mRNA purification kit (Invitrogen). After mRNA fragmentation, cDNA synthesis was primed with random hexamers and libraries constructed with the mRNA-seq sample prep kit (Illumina, San Diego, CA) according to manufacturer's recommendations. The cDNA libraries made from leaf base and tip were sequenced on an Illumina Genome Analyzer 2. The sequencing reads were aligned to the maize genome AGP and a splice junction database constructed with Eland software (Illumina), allowing up to 2 mismatches in the alignments.

On the basis of the annotated coordinates, each read was assigned to different categories of exon, intron, intergenic, splice junction and repeats. The reads that were aligned to exons were used to calculate the transcription level of each gene model. The background level is estimated by calculating the alignment density in genomic regions at least 10kb away from any annotated gene boundaries. The gene models were considered actively transcribed if the depth of coverage was found to be significantly above the background level when allowing a false-positive rate of 1% (assuming a Poisson distribution of coverage depth at each base).

Transcriptomics of shoot apical meristems (*SAMs) and seedling from reciprocal hybrids*

SAMs from 14 DAP B73 seedlings were harvested, fixed, embedded in paraffin, sectioned and tissues collected via laser capture microdissection. L1 and L2 layer tissues were collected separately from the same SAMs. In total, a pool of RNA sampled from the L1 of 13 SAMs and a pool of RNA samples from the L2 of 13 SAMs were extracted followed by RNA amplification and synthesis of double-stranded cDNA (*S23*).

RNAs were extracted from a single replication of 14-day-old B73xMo17 and Mo17xB73 seedlings. For the purposes of this experiment data from the two reciprocal hybrids were computationally merged. RNAs were purified with DNaseI treatment followed by cleanup with the RNeasy Plant Mini Kit (Qiagen, Valencia, CA) as per manufacturer instructions. Sequencing library construction was completed with the Illumina mRNA-Seq sample preparation kit.

Both RNA-seq reads from SAMs and reciprocal hybrids were aligned to maize gene models (http://maizesequence.org) with short read aligner NOVOALIGN (http://www.novocraft.com). Two mismatches across 32 bp were allowed and only the reads that uniquely mapped to gene models were considered for further analysis. RNA-seq reads have been deposited in the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo). Accession numbers for the RNA-seq experiments are: GSE16136 (reciprocal hybrids), GSE16868 (L1-L2 of SAM) and GSE16916 (seedling leaves).

Results

To validate gene expression data, RNA-seq analysis was performed on cDNA synthesized from RNA isolated from the base and tip of a 9 day-old Mo17 seedling leaf. To broadly survey gene expression in the leaf, we examined expression in sink (actively importing sugars) and source (photosynthetically active) tissues of a third leaf as it was emerging from the whorl (see Methods above). Approximately 48 million reads were generated from base cDNA samples and 46 million reads from tip samples. Among the 32,540 annotated genes (53,764 transcripts), transcriptional activity was verified for 27,222 genes (47,693 transcripts) in the leaf base sample and 26,336 genes (46,313 transcripts) in the leaf tip sample for a total of 28,560 genes.

Approximately 80% of the Illumina reads mapped to annotated exon sequences and 5% to exon junctions (see Fig. S10). The 8.4% of reads mapping to intergenic or intron sequences may represent expressed non-coding RNA's or misannotated exon sequence. Approximately 7% of sequence reads map to repeat masked regions of the AGP, likely representing actively transcribed transposon sequences. The leaf transcriptomics data also provides support for over 80% of the FGENESH models that did not have EST support. The relatively high support of FGENESH models in the AGP with Illumina data suggests a low false positive rate for FGENESH models.
RNA-seq experiments were also conducted with RNA samples extracted from 14 day old B73 shoot apical meristems (SAM; L1 and L2 cell layer, collected separately but computationally pooled for the analyses reported here) and RNA samples from 14 day old seedlings from two reciprocal hybrids (B73xMo17 and Mo17XB73). Among the filtered gene set, transcription support (defined as having at least 2 RNA-seq reads) was obtained for 61% (20,011/32,540) and 81% (26,394/32,540) of the genes from the B73 SAM (~14 million reads) and reciprocal hybrid (~15 million reads) RNA-seq experiments, respectively. As shown in Fig. S11, the pooled data from the three RNA-seq experiments provide evidence for the transcription of 91% of the genes in the filtered gene set (29,541/32,540).

Centromere Methods

Identification and draft sequencing of centromeric BACs

Centromeric BACs were selected for draft sequencing with data from the maize FingerPrinted Contigs (FPC) project obtained from the Arizona Genomics Institute and dated Oct 27th, 2004 (*S41*). Presumed centromeric FPC contigs were identified on the basis of containing a large number or percentage of BACs with centromeric hybridization or BAC end sequence homology to centromere repeats CentC and CRM. Subsequently, BACs that had CentC homology and did not assemble into FPC contigs (singletons) were also sequenced. The 101 selected BACs were shotgun sequenced at the Clemson University Genomics Institute by Chris Saski, resulting in 27,936 reads. Following cross_match with vector sequence and the *E. coli* genome, 14,911,251 high quality nucleotides remained. Sequences were deposited into the NCBI Trace Archive with TI numbers 1757396377-1757412600 and 2185189231-2185200942 and used to generate centromeric markers (*S41*).

Enrichment and representation calculations

A total of 1,087,012,190 high quality nucleotides (nt) from 1,124,441 whole genome shotgun (WGS) reads generated by the Joint Genome Institute were downloaded from NCBI's Trace Archive on Feb 22nd, 2008 (SPECIES_CODE = "ZEA MAYS" and CENTER_NAME = "JGI"). Vector sequences in these reads were masked with cross_match (http://www.phrap.org/) and NCBI's UniVec_Core build #5.1 from Jan 13th, 2009. Tandem repeat homologies were calculated with cross_match. Overlapping nucleotides were assigned to the repeat with the longer contiguous sequence homology. Nucleotides homologous to retrotransposons were calculated by competitive WU-BLAST (http://blast.wustl.edu/) with score filters S/S2=1000 for the B73 RefGen_v1 and S/S2=225 for WGS and ChIP datasets. All HSPs with identity ≥90% were collected. Only the longest HSP was used to assign sequence homology to a specific subject database when HSPs overlapped. Gene sequences were identified with WU-BLAST with a score filter of S/S2=225 for all genomic datasets and limited to ≥90% identity HSPs. Databases used include a set of 79 CentC records (AF078918- AF078923; AY321491; AY530216- AY530287), Cent4 (AF242891.1), knob-180 (M35408.1), and knob-350 (AF071121.1-AF071124.1) sequences obtained through GenBank, as well as 5S and 45S rDNA consensus sequences that were created from BAC (AC208721 for 5S rDNA) and WGS sequences. The presumed 45S rDNA sequence was reconstructed from the 100 GenBank *Zea mays* WGS traces with the highest BLAST scores to X03990 and their mate pairs – contact Gernot Presting (gernot@hawaii.edu) for sequences. For the rDNA, Cent4, knob-180, and knob-350 databases, two copies of each sequence were concatenated in tandem in order to capture short sequence homologies covering repeat junctions. A set of all previously identified CRM1, CRM2, CRM3, and

CRM4 elements (*S42*) was used in competitive BLAST experiments. The opie database consisted of 70 different full-length elements extracted from BACs (*S43*). The gene set was composed of 1,266 mRNA sequences obtained after removing transposase, chloroplast and mitochondrial sequences from the "*Zea mays*" portion of RefSeq downloaded on Feb 4th, 2008. Results are summarized in Table S13.

Mapping ChIP sequences to BACs or B73 RefGen_v1

MUMmer. The 16,875 BACs (downloaded from GenBank htgs on Feb 23rd, 2009 and comprise 2,833,162,209 non-N nt) were split into sets totaling no more than 100 Mb. Each of the resulting 10 sets, plus the chloroplast (NC_001666.2) genome were used as a database in a MUMmer search with 149,756 sequences obtained from chromatin immunoprecipitated with anti-CENH3 antibody (*S44*). All perfect matches over the entire length of a ChIP sequence in either direction were recorded. Reads that matched one or more locations on no more than two sequences (BAC or chloroplast genome) were mapped. This was done to account for overlapping BACs in the minimum tiling path and overlapping sequence contigs within the same BAC. BACs with more than 15 MUMmer hits were classified as "hi-ChIP". Alternatively, single reference chromosomes were used instead of BAC sets. Because most overlap was removed in the B73 RefGen_v1, all reads that matched a single location in only one reference chromosome and did not match the organellar genomes (44,897) were mapped.

BLAST. The 11 reference chromosomes and the chloroplast genome were formatted into a database that was searched with all ChIP reads (NCBI BLASTN -F F, -W 9, -e 1e-3). Reads (104,810) were mapped to the genome region with the highest bitscore if the HSP had a minimum 96% identity over 96% query length and provided that all other HSPs had a lower bitscore.

Identification and mapping of centromeric EST and cDNA sequences

A BLAST database was created from 72,026 cDNA and 2,018,338 EST sequences of *Zea mays* downloaded from GenBank on Mar 29th 2009 and searched (NCBI BLASTN, e = 1E-20) with consensus sequences for CRM1, CRM2, CRM3, CentA, CentC, as well as 106 high ChIP BACs previously masked for CRMs, CentC, knob repeat as well as mitochondrial and chloroplast genes with crossmatch. False hits due to non-centromeric homologs were eliminated by BLAST of the unique cDNA and EST sequences identified with the high ChIP BACs against a database of 16,875 *Zea mays* BACs. A total of 3,075 non-CRM/CentC transcripts (3,002 ESTs and 73 cDNAs) were mapped to the functional centromeres by reciprocal BLAST. BLASTN was used to map 396 of the 398 B73 transcripts to the B73 reference chromosomes. BLAST output was manually inspected to ensure that HSPs covered most of the transcript length and that each transcript was mapped to a single locus or at most two locations separated by no more than 7 kb.

Transcripts that mapped to the same 16 kb window were considered to have originated from the same locus. The loci were assigned to genes with the 4a.53 filtered and non-filtered gene sets. EST library information was obtained from (*S45*) and the library browser (http://www.ncbi.nlm.nih.gov/UniGene/lbrowse2.cgi?TAXID=4577) at GenBank.

Synteny Analysis and Identification of the Maize Lineage Whole Genome Duplication

Methods

The genomes of rice and sorghum were used as references to demarcate the whole genome duplication event that occurred specifically in the lineage leading to maize. In particular, the 10-chromosome configuration of sorghum closely resembles the ancestral state of maize's subgenomes (*S1*). Conserved synteny between maize-sorghum, and maize-rice was investigated with respective sets of phylogenetically-determined ortholog pairs. For maize-sorghum there were 27,275 ortholog pairs consisting of 25,216 maize and 20,408 sorghum genes. The maize-rice set included 31,586 pairs between 25,844 maize and 20,569 rice genes. In total 27,550 maize genes were included in this analysis. First, DAGchainer (*S46*) was used to identify colinear chains. To increase sensitivity, chaining was created on the basis of gene order, excluding positions of non-orthologous genes, rather than with gene coordinates. Chains were required to have at least five colinear genes with no more than ten intervening genes between neighbors. Resulting gene-pairs were classified as "syntenic:colinear". Next, we searched for additional syntenies among non-colinear genes to account for small-scale rearrangements and assembly errors. This also allowed members of tandemly duplicated clusters to be classified as syntenic, as they were often missed when strict colinearity was enforced. For a non-colinear gene in species1, nearest flanking colinear anchors were identified and mapped to their respective positions in species2. If the non-colinear ortholog in species2 was located within ten genes of either anchor, then the non-colinear pair of genes was classified as "syntenic:in-range". Nearly all of these occurred within already defined colinear chains. Calculations of synteny coverage were from the collapsed (i.e. non-overlapping) coordinates of colinear chains. Alignments between grasses typically reveal ancient duplication events that occurred in the progenitor of grasses (*S1, S29, S47*). Such paralogous relationships were screened out by looking for 2:1 correspondences between the rice or sorghum references and maize. To define duplicated regions that arose from the maize-specific whole genome duplication we looked for "co-syntenic" maize genes, i.e., those residing on different chromosomes but syntenic to the same sorghum or rice reference gene.

Results

Fig. S14 shows dot plots resulting from alignments between orthologous gene sets of maize-sorghum and maize-rice. As expected, most regions of sorghum and rice have a 1:2 relationship with maize, indicative of the genome-wide duplication in maize. Use of orthologs precluded alignments between remnants of the ancient polyploidy event common to grasses (*S1, S47*). However, undesired paralogous relationships were detected over regions of the ancient segmental duplication found on rice chromosomes 11 and 12, and corresponding regions of sorghum chromosomes 5 and 8 (*S29*). Frequent gene conversions between these regions have resulted in paralog homogenization (*S48*), which contributed to the small numbers of false-positive ortholog calls observed here. Such relationships, specifically between maize chromosomes 10 and 3 with sorghum 5 and rice 11, and between maize 2 and 4 with sorghum 8 and rice 12) were prior to additional analyses.

Overall, a total of 23,589 maize genes had syntenic relationships, representing ~86% of the starting set of orthologous genes (see Table S14) and covering ~89% of the maize genome (1,832 Mb/2,061 Mb). Of these genes, 4,010 were syntenic exclusively to sorghum, whereas 1823 showed synteny only to rice. This translated into ~86% of maize-sorghum orthologs (within the maize-sorghum set) having synteny with sorghum compared to ~76% with rice. Conversely, ~90% of sorghum orthologs were syntenic to maize compared to ~77% of rice. This suggests that greater numbers of gene movements have occurred since the maize-rice lineage split as compared to that maize-sorghum, as expected given maize's closer relationship to sorghum.

Co-synteny of unlinked maize genes to a common rice or sorghum reference gene gave 8110 retained duplicate homoeologs, representing ~25% of the identified maize genes (8,110/32,450) and ~29% of orthologous genes (8,110/27,550). This number is expected to be somewhat of an underestimate because some retained duplicates would fail to have a co-syntenic relationship in cases where either of the maize genes moved to a new location or where the reference gene in rice or sorghum suffered lineage-specific gene losses or movement. However, the alternatives of our using all maize paralogs or co-orthologs to perform a maize-to-maize alignment produced high levels of noise when DAGchainer was used. The maize-maize dot-plot in Fig. S15 reveals duplicate regions on the basis of retained homoeolog pairs.

Preferential Retention of Gene Functional Classes

Enrichment for functional classes in retained duplicate homoeologs and in singletons was tested with the GO Molecular Function and GO Biological Process terms (*S49*). This analysis used a hypergeometric test with the Benjamini-Hochberg correction (*S50*) as implemented with BiNGO software (*S51*). Because each GO term inherits all annotations from its descendents, there was considerable overlap of genes between statistically significant terms. Results are shown in Table S15.

Preferential Gene Loss in Ancestral Homoeologous Chromosomes

Syntenic blocks were assigned to ancestral homoeologous chromosomes with sorghum used as a reference, which approximates ancestral structure of the maize genome prior to its whole genome duplication event. Each syntenic sorghum gene was classified as being: a) retained singly at one homoeologous site in maize; b) retained singly at the other homoeologous site; or c) retained at both sites. Results were tallied across each sorghum chromosome, as shown in Tables S16, S17 and Fig. S17.

Identification of Paralog Clusters

Tandem clusters were defined as paralogous genes positioned with no more than two intervening genes. Members of tandem clusters could be retained at homoeologous positions. Table S18 summarizes the largest tandem clusters on the basis of the sum of cluster sizes across homoeologous sites.

Maize Finishing Guidelines

Non-repetitive portions of the sequence have had sequence improvement (directed attempts) and have been labeled as "improved." Improved regions are double stranded, sequenced with an alternate chemistry or covered by high quality data (i.e. phred quality greater than or equal to 30 or approval by an experienced finisher), unless otherwise noted. Regions of low sequence complexity (such as dinucleotide repeats and small unit tandem repeats) in the improved regions have not been resolved to previously established finishing standards. BAC end sequence, C*ot* and methyl filtered Genome Survey Sequence and data from overlapping projects of strain B73 may have been included in this project. Order and orientation of contigs will be performed after the finishing work has been completed.

Rules for determining regions to be finished in the maize genome:

After BACs have received approximately 4-5X coverage, they will be assessed to determine which areas will receive improvement. The projects will be compared to a repeat database to eliminate maize repeat sequence from the regions to be tagged for improvement. The projects also will be screened for regions to be improved. The in-house repeat tagging script should be rerun each time that new data is added to an assembly.

a) Regions then will be tagged for finishers to improve on the basis of the following criteria:

a. Existing Contigs of 1kb or larger in size with greater than 4 subclones:
Regions of good quality unique data of 500 bases or more will be improved.
Good quality is defined as majority phred20 or higher and will be tagged.
500 base pairs of repeat flanking the region will also be improved on each side, within the confines of the BAC. It is assumed that 500bps of contiguous sequence will be adequate to locate regions of interest to improve.

b. Gaps:
If gaps are flanked by 1kb of repeat sequence, no attempts will be made, assuming that the gap is more repeat, unless the gap is contained in a region of significant similarity to known gene sequence. Any unique regions of high quality data (phred20) of 400bps or greater on either side of a gap must have two attempts to be resolved unless there is no information to order and orient the gap.

c. Extra Contigs:
E. coli contamination and data determined to be contamination will be removed from the databases. Data with a BLAST hit to a different organism to be considered contamination. Small contigs of less than 1kb in size or 5 subclones and with less 1000 bps of unique sequence will not be improved unless they are recognized as regions of significant similarity to known gene sequence or unique sequence. Small contigs

that are 1kb or smaller in size or less than 5 reads and are not included in the final assembly will be deposited in a database with BLAST capacity. As part of post-processing, doFinish tags (in-house program) are removed from contigs that are less than 1000 bases and from regions within contigs that are less than 500 bases, except at gaps when they are removed from regions less than 400 bases.

b) BAC ends:

We will attempt to identify the BAC end/vector junctions of all projects to be assured that unique or significant data will not be overlooked. The BAC end/vector juncture should be identified, even if it falls within repeat sequence.

c) Entire Projects Tagged as doFinish:

Projects tagged entirely as doFinish should be submitted to the BLAST program to verify that they are truly of maize origin. Projects that have strong BLAST hits to other organisms and do not have any C*ot* and methyl filtered Genome Survey Sequence data should be investigated as possible contamination or mix-ups with the library and determined to be of maize origin before submission as improved sequence.

Guidelines for improvement of the maize genome:

Transposon Sequence

a) Transposon sequence should be identified and excised from submitted sequence, even if it does not fall within a doFinish region.

Misassemblies

a) If a read pair has one mate in unique sequence and one mate pair in a repeat region, the project should be assembled, so that the read pairs correlate correctly. Misassemblies in repeat only regions do not need to be sorted. Autoedit (Gordon, http://www.phrap.org/consed/consed.html) will be applied to separate piled repeats.

Ambiguous Bases

a) Attempt to resolve ambiguous bases with a reaction from each direction, most likely utilizing a 4:1 chemistry walk on a subclone. If the subclone walk is unsuccessful, attempt PCR and sequence from both directions to resolve. If the region remains unresolved, annotate the region as "non-repetitive but unresolved" and add a comment tag with the attempts listed for QA purposes.

Data from confirmed map neighbors used to improve BAC clones

a) Data from overlapping projects should be used to improve projects as long as they are confirmed map neighbors. Confirmed neighbors should not have any polymorphisms or base differences other than in mononucleotide runs and simple sequence repeats. An annotation tag is needed if the data is from libraries other than B73. Neighbor data should not contain any high quality discrepancies. The program getoverlaps can be used to list verified neighbors and ~akozlowi/bin/maize_TP clonename for newly entered data.
b) C*ot* and methyl filtered Genome Survey Sequence data or mRNA data should be utilized when it can improve the assembly. Improved finish regions with only C*ot* and methyl filtered Genome Survey Sequence data or mRNA data should be annotated as "GSS and/or mRNA data only."
c) Regions of Genome Survey Sequence data or mRNA data only should have a minimum of 100bps overlap with BAC data and all overlapping data should have no more than two true base differences.
d) Contigs greater than 2kb in size that contain only gss and/or gll data should be removed from the database. If they are smaller than 2kb they may remain in the database; if they are not in a scaffold, any doFinish tags should be removed.
e) Fosmid end walks should be added to projects to verify sequence and to help orient contigs. The program get_maizeFES (local program) is run in prefinish with the -nomate option, but is recommend that finishers

use the option that will pull in fosmid mate pairs, get_maizeFES –clone and then remove fosmids which do not belong.

f) Unresolved dinucleotide runs

For dinucleotide repeats where the repeat unit is similar on both sides of the gap, no additional closure attempts are necessary except for an annotation tag.

Gaps in Unique regions

a) Gaps that are spanned and oriented should have subclone walks and PCR walks attempted from both directions to resolve the gap before annotating as "non-repetitive but unresolved region." If the gap remains after 2 rounds of reactions and still has doFinish tags on an end, it needs to be assessed to determine if further reactions would be beneficial. If the previous reactions were successful and did not end in repeat or sequence that would be difficult to extend, more subclone walks and/or PCR should be attempted. If the new data did result in repeat or structure that a subclone walk or PCR walk would not likely add data, then the region should be tagged as "non-repetitive but unresolved" and should be comment tagged with the attempts made for QA purposes.

b) Gaps that are oriented but not spanned should have 2 attempts at PCR before annotating as "non-repetitive but unresolved." If the region remains unresolved, then it should be comment tagged with the attempts made for QA purposes. Gaps that are not spanned and not oriented will have no finishing attempts made and should be tagged as "non-repetitive but unresolved region."

c) Low quality data on the ends of contigs can be replaced by Ns and the doFinish tag removed from them, if they do not seem to represent the true sequence. If in the finishers estimate the sequence is most likely not base perfect but may provide some insight into the correct sequence of the gap, it should remain but be tagged as "non-repetitive but unresolved region." Data should not be clipped just for one unresolved base or for a dye blob that normally could be easily resolved. When a gap is closed after making appropriate attempts but bases remain unresolved, additional attempts to resolve the bases should be made, if it is likely that the bases could be resolved with further walks. An example would be that bases remain unresolved because the sequencing reactions are at position 700 or greater where the data can be poor because of the length of the run and not due to any structure or repeat sequence.

PCR

a) PCR attempts should be made with cloned DNA, not genomic DNA (why?).

BAC end sequence

a) Both cloning sites should be identified in a project and will be used after submission to verify that the data represents the correct BAC. The program to retrieve BAC ends for maize projects is get_maize_BES. The program will prompt the user for a project name. A finisher can assemble the vector to help find subclone templates to sequence to find the BAC end. If that is unsuccessful, two attempts at PCR should be made. If the region remains unresolved then it should be comment tagged with the attempts made for QA purposes.

Projects with no doFinish regions

a) Projects that have no doFinish tags should have the program tag_maize (local program) re-run prior to presubmitting to verify that there are no regions to improve (no doFinish tags). A comment tag should be made in the "other comments" stating that there are no doFinish regions in the project and that the finisher has verified that the tag_maize program has run correctly. Coordinator approval is not necessary. If the tagging program tags regions as doFinish that do not meet the guidelines (i.e., too few reads or below phred20), you can remove the doFinish tags after leaving proper tags (comment or coordinator approval). Projects that have no doFinish tags should contain repeat tags and gll or fosmid reads.

Reassembling and Retagging

a) All projects should have new data entered with the assembly program, Phrap, which will reassemble and tag the doFinish regions. There will be some exceptions to this rule for projects that are extremely repetitive and have had extensive sorting done to them that reassembling would waste. These should have notes in the notes file explaining the situation. When reassembling cannot be done, tag_maize should still be run to verify the doFinish regions. Reassembling is preferred in most cases, because the new data may have matches to singlets that were not in the original assembly and may help to resolve the doFinish regions.

b) Tag_maize should be run again on projects after extensive editing has been done, in case these changes would increase doFinish regions, and prior to presubmit, so that in the event of updates to the tag_maize prthogram the most current information is submitted.

Contamination in maize projects

a) If there are many small contigs with no forward-reverse linkage to the main contigs or scaffolds, report that to the Finishing Manager, who is investigating contamination in maize projects so that improper data is not submitted with maize projects.

More data needed

a) After checking overlaps, if there are more than 5 or 6 scaffolds and low coverage overall, a project may need more data. Please forward requests for more data to a Finishing Manager to address.

Computer programs used to aid maize finishing (local programs)

a) tag_maize --a --q –nav: tags the regions in the project targeted for improvement and produces a consed navigation file of the tagged regions.

b) gss_gll_nav: produces a consed navigation file pointing to regions that are covered solely by high $C_o t$ methyl filtration reads (GSS) and/or mRNA reads (GLL).

c) supernav -a <acefile> --navs <cafcop> --lists list.txt,list.txt2 –o <outputfile> --pt: combines several different consed navigation files in a single file, which allows for more efficient navigation of the project by the finisher

Supporting Figures

Figure S1. Distribution of genes (exons) and transposable element repeat classes across the ZmB73v1 assembly. Base-pair percent-coverage was calculated in 1 Mb sliding windows, incrementing every 200 kb. Track "Non-CACTA DNA TE" includes Class II elements Mutator, hAT, Pif-Harbinger and Tc1-Mariner. Arrow indicates the position of centromeres. Scale (min = blue; max = red). See Table S4 for the range of minimum and maximum values for each track.

Figure S2. Process for selecting the filtered gene set on AGP. The projected genes are from direct coordinate mapping between intra-BAC contigs and the AGP. The mapped genes are those predicted on BAC contigs that are not used in the AGP; they are aligned to the AGP on the basis of sequence similarity. See supporting online text for details on the filtering process.

Figure S3. Species comparison of exon length distribution. Arabi-zm-ort: indicates Arabidopsis genes with orthology to maize; rice-zm-ort indicates rice genes with orthology to maize; sorghum-zm-ort indicates sorghum genes with orthology to maize; maize-all indicates all maize genes in the filtered set; maize ort indicates maize gene with orthology to Arabidopsis, rice or sorghum genes; and maize-cdna: indicates maize genes built with pure fl-cDNAs.

Figure S4. Species comparisons of intron length distributions. This cumulative distribution plot shows fraction of introns with lengths exceeding threshold. See Fig. S3 legend for explanation of gene sets used in this analysis.

Figure S5. Exon GC percentage. See Fig. S3 legend for explanation of gene sets used in this analysis.

Figure S6. Intron GC Percentage. See Fig. S3 legend for explanation of gene sets used in this analysis.

Figure S7. Transcript GC percentage. See Fig. S3 legend for explanation of gene sets used in this analysis.

Figure S8. Comparison of protein length distributions among species annotations. See Fig. S3 legend for explanation of gene sets used in this analysis.

Figure S9. The 10 most common Gene Ontology (GO) terms in the Filtered Gene Set contains for each functional class.

Figure S10. Distribution of RNA-seq reads from seedling leaf to the ZmB73v1. Approximate number of 32 nt reads from seedling leaves are shown (in millions) together with percentage of placed reads to each class. No placement data was obtained for approximately 22% of reads. Of the 32,540 genes in the filtered gene set, 26,857 evidence-based and 1,703 FGENESH models were verified. Percentages do not sum to 100 due to rounding.

Figure S11. Number of genes from the filtered gene set (N= 32,540) whose transcription is supported by RNA-seq experiments. In combination the three RNA-seq experiments provide evidence for the transcription of 91% of the genes in the filtered gene set (29,541/32,540).

Figure S12. Centromere repeats of maize chromosomes. This composite image of chromosomes 1, 2, 3, 8 and 9 shows the position of the centromere on the basis of anti-CENH3 ChIP and the location of CRM and CentC repeats. CENH3 density is displayed as a moving average of the sum of the number of 454 reads per 0. 1% chromosome arm length for all 5 chromosomes, calculated over nine windows. Blue line = uniquely mapped

reads at 100% identity, red line = single best match at ≥96% identity and ≥96% read length. CentC and CRM repeats are shown as the sum of the number of nucleotides per 0.1% chromosome arm length for all 5 chromosomes. Note the exponential y-axis for the CentC panel due to the high concentration of CentC in the centromeres and the variable scales for the CRM panels. Arrow = centromere.

Figure S13. Distribution of DNA Methylation and Heterochromatin in Maize and Sorghum. A) Approximately 567,000 B73 methyl filtration reads were retrieved from Genbank (*S52*) and aligned to the ZmB73v1 assembly (CP) with BLAST. Chromosomes were divided into 0.1 Mb bins and the number of reads in each bin that aligned uniquely in the sequenced genome with at least 95% identity were counted. Per-bin counts were divided by the total number of reads mapping uniquely in the genome to yield relative enrichment or depletion scores for coverage in each bin. Red tones reflect enrichment in MF sequence density, while green tones reflect depletion. B) 534,000 sorghum bicolor reads (*S53*) were retrieved from Genbank and aligned to the SbiI sorghum genome (*S29*) with BLAST. Relative enrichment or depletion scores for coverage across 0.1 Mb bins was calculated as in maize. C) The MF density map for B73 chromosome 8 is compared to DAPI, CENPC, and H3K27me3 staining performed by Shi and Dawe (*S54*).

Figure S14. Dot-plots showing maize-rice and maize-sorghum orthologs, classified as not-syntenic (black), syntenic:colinear (blue), and syntenic:in-range (red). See supporting online text for explanation of these terms.

Figure S15. Maize versus maize dot-plot. Black dots show “co-orthologs”, maize genes that have orthology to the same rice or sorghum gene. Red dots show “co-syntenic” maize genes, those that are syntenic to the same rice or sorghum gene.

Figure S16. Genes of the CesA/Csl superfamily. At least three distinct cellulose synthase (CesA) genes are co-expressed during primary wall formation and secondary wall formation; mutants in each of them result in cellulose deficiencies, indicating that all three are essential for cellulose synthesis (*S55*). Rice, maize, and sorghum have genes in apparent paralogous clusters with the Arabidopsis CesA genes. Whereas Arabidopsis and rice have ten CesA genes, four additional duplications have occurred in sorghum, and ten in maize. The functions of the cellulose synthase-like (Csl) genes are beginning to be established. The CslA genes are associated with b-mannan synthesis (*S56, S57*), CslC genes are involved in xyloglucan formation (*S58*), and some of the CslD genes may be involved in cellulose synthesis in tip-growing cells (*S59, S60*). The CslF subfamily of genes is found only in grass species, and heterologous expression indicates its involvement in the synthesis of the grass-specific mixed-linkage (1→3),(1→4)-b-D-glucan (*S61*). In contrast to the CesA family, maize has retained fewer recent duplications in the CslFs than has sorghum. The CslD family comprises five paralogous clusters of each of the grass species, with no recent duplications. The CslA family is greatly expanded in the grasses, with evidence of nine paralogous duplications in grass specific subclades. Grass-specific duplications sometimes resulting in expansion to create new subfamilies is common among the grasses (*S62*). For accession numbers of all genes in this superfamily and other cell wall-related gene families, see http://cellwall.genomics.purdue.edu/families/.

Figure S17. Preferential gene loss between homoeologous maize chromosomes. Percentages of syntenic sorghum genes whose maize syntenic ortholog is present as singletons or retained at both homoeologous sites are shown. Scale is gene index.

Figure S18. Distribution of NIP/TIP pairs. 222 NIPs exhibit 100% identity; these TIPs (Totally Identical Paralogs) are highlighted in red. (A) Local NIP/ TIP pairs are located within 200 kb of each other; (B) Distributed NIP pairs are >200 kb apart or on different chromosomes. It has been hypothesized that different mechanisms are responsible for the origins of these two classes of NIPs (*S63*). Although NIPs are distributed across the genome, some regions (e.g., 2L and 4S) have elevated rates of inter-chromosomal NIPs. These do not, however, reflect known segment duplication events, arguing against paralog homogenization as a mechanism for the origin of NIPs. On the basis of aCGH experiments ~5% of NIPs have stronger signals in B73 than Mo17

genomic DNA (*S64*), suggesting that Mo17 may have only a single copy of what in the B73 genome are NIP pairs. Centromere positions are from (*S44*).

Supporting Tables

Table S1. Estimation of maximum rate of sequencing errors. The maximum rate of sequencing errors was estimated with as a gold standard 10 genes (start to stop) cloned from B73 and sequenced by the Schnable Lab. The sequences of each gene had been carefully and manually edited and subsequently aligned to MF and HC reads (*S1*) during which process a few remaining errors were corrected. Eight of the ten genes aligned almost perfectly to the B73 reference genome. These genes, comprising 32 kb of aligned sequence were used to estimate the maximum rate of sequencing errors.

Table S2. Transposable elements (TEs) in the B73 RefGen_v1

Table S3. TE-related InterPro IDs used for exclusion of proteins from the Filtered Gene Set. Protein sequences were annotated for InterPro domains as described above. InterPro domains were chosen on the basis of association with independently-classified TE's and screened to ensure specificity.

Table S4. Minimum, maximum and average base-pair % coverage for each panel shown in Fig. S1. Values represent base-pair percent-coverage calculated in 1 Mb sliding windows, incrementing every 200 kb.

Table S5. Comparison of orthologous genes in maize, rice, sorghum, and Arabidopsis. Gene orthologs were defined phylogenetically with EnsemblCompara GeneTree method as described (*S30*). Sources of annotation: the maize filtered set; the sorghum genome (*S29*) JGI release Sbi 1.4 from March 2008, The Arabidopsis Information Resource (*S31*) release 8 from April 2008, and the MSU/TIGR Rice Genome Annotation Resource (*S28*) release 5 from January 2007.

Table S6. Summary of introns with repeats. The RefGen_v1 sequence was masked with the TE Consortium repeat library. A subset of genes (21,491) built purely from FLcDNAs were examined. Introns and genes that contain repeats were counted and percentages given.

Table S7. Comparison of annotations generated by exclusive use FLcDNA versus the maize filtered set and other annotation sets.

Table S8. Quality assessment of 16 predicted starch-pathway genes in the filtered set. Established gene-structures from the maize starch biosynthetic pathway were chosen to assess the quality of the annotations in the maize filtered set.

Table S9. Validation of gene structures of filtered gene set (FGS). Coding sequences of 10 gold standard genes were aligned to the B73 reference genome at ≥96% identity and 100% coverage. All filtered genes found within the same intra-BAC contig were projected onto the original BAC sequence from which the contig was derived. Seven of the ten genes have complete coverage in the filtered gene set. Only a partial gene models overlaps with the last exon of *rf2d* and no gene model was found for *pdc3*. Both genes have EST and FLcDNA support, do not contain known repeats within the annotated gene models and are located in the middle of the intra-BAC contigs. Two genes (4 coding sequences) map to *rf2a*, however, intron 5 was skipped in the model. The ~7kb intron 5 of *rf2a* contains retrotransposon sequences. It is conceivable that the associated gene models were truncated for reasons of size and/or repeat content. However FLcDNAs for this gene are available in Genbank as indicated in the body of the table.

Table S10. Correspondence between coding sequences (CDS) known from full-length cDNA clones and computational predictions from the maize genome sequence. Enzymes of the starch biosynthesis pathway were analyzed. cDNAs are identified by Genbank accession numbers (gi) and by genetically defined loci in those instances where mutations have been identified. “First nt.” and “Last nt.” indicate the starting and ending positions of the predicted gene in the chromosome pseudomolecule. Numbers in the correspondence description refer to the known coding sequence inferred from the cDNA.

Table S11. Top 50 InterPro hits in the Filtered Gene Set. Shown are counts and percents of genes bearing the listed domain.

Table S12. Possible sources of artifacts that explain rice-sorghum specific families. Peptide queries from rice-sorghum specific families were aligned to the indicated databases with TBLASTN (Wu-blast, P-value cutoff 1e-20). In total there were 108 non-TE rice genes and 96 non-TE sorghum genes that aligned to maize cDNA, but not to the maize genome sequence; indicating genes not annotated due to missing genomic sequence.

Table S13. Representation of tandem repeats and retrotransposons in the B73 RefGen_v1 and their enrichment in the ChIP reads. The numbers of nucleotides attributable to each tandem repeat in the whole genome shotgun (WGS) data, the B73 reference genome, and ChIP reads were determined with competitive cross_match. Competitive WU-BLAST was used to quantify CRM and opie elements, as well as genes, in each data set. Note that the CRM elements are well represented in the reference genome. Sequence representation in B73 RefGen_v1, and enrichment in the ChIP fraction were calculated for each data set by dividing the fraction of nucleotides represented by a sequence in B73 RefGen_v1 or the ChIP sequence by the fraction of nucleotides represented by that sequence in the WGS data.

Table S14. Counts of genes showing synteny in comparisons of maize to sorghum and maize to rice, and their percentage relative to orthologous genes used as input for this analysis.

Table S15. Enrichment for functional classes in retained duplicate homoeologues. Enrichment for functional classes in retained duplicate homoeologs and in singletons was tested with the GO Molecular Function and GO Biological Process terms (*S49*). Those with significant P-values are shown.

Table S16. Classification of sorghum loci with respect to syntenic positions in maize. Sorghum was used as a reference to assign syntenic maize genes to their ancestral homoeologous chromosomes, A and B. Each sorghum gene was then classified with respect to the disposition of corresponding sites in maize, either present as a singleton in A, singleton in B, or as duplicates retained in both A and B.

Table S17. Raw data summarized in Table S16. This table shows syntenic gene counts broken out by maize chromosome components.

Table S18. Top thirty paralog clusters and their distribution across homoeologous sister regions. Clusters were defined as proximally located paralogs that have no more than two intervening non-paralogous genes. Clusters were ranked on the basis of the sum of cluster sizes at both homoeologous sites.

Supplementary References

S1. F. Wei *et al.*, *Plos Genet* **3**, e123 (2007).
S2. F. Wei *et al.*, *Plos Genet*, submitted.
S3. Z. R. Bao, S. R. Eddy, *Genome Res* **12**, 1269-1276 (2002).
S4. R. Bruggmann *et al.*, *Genome Res* **16**, 1241-1251 (2006).
S5. G. Haberer *et al.*, *Plant Physiol* **139**, 1612-1624 (2005).
S6. B. A. Kronmiller, R. P. Wise, *Plant Physiol* **151**, 483-495 (2009).
S7. J. Lai, Y. Li, J. Messing, H. K. Dooner, *Proc Natl Acad Sci U S A* **102**, 9068-9073 (2005).
S8. J. Lai *et al.*, *Genome Res* **14**, 1924-1931 (2004).
S9. R. J. Langham *et al.*, *Genetics* **166**, 935-945 (2004).
S10. W. Ramakrishna, J. Emberton, M. Ogden, P. SanMiguel, J. L. Bennetzen, *Plant Cell* **14**, 3213-3223 (2002).
S11. R. Song, J. Messing, *Plant Physiol* **130**, 1626-1635 (2002).
S12. Z. Swigonova *et al.*, *Genome Res* **14**, 1916-1923 (2004).
S13. A. L. Rayburn, D. P. Biradar, D. G. Bullock, L. M. Mcmurphy, *Heredity* **70**, 294-300 (1993).
S14. Y. Han, J. M. Burnette, 3rd, S. R. Wessler, *Nucleic Acids Res* **37**, e78 (2009).
S15. R. C. Edgar, *BMC Bioinformatics* **5**, 113 (2004).
S16. D. L. Swofford, (Sinauer Associates, Sunderland, MA, 1999).
S17. K. Tamura, J. Dudley, M. Nei, S. Kumar, *Mol Biol Evol* **24**, 1596-1599 (2007).
S18. L. Yang, J. L. Bennetzen, *Proc. Natl. Acad. Sci. USA*, in press.
S19. L. Yang, J. L. Bennetzen, *Proc. Natl. Acad. Sci. USA*, submitted.
S20. R. S. Baucom *et al.*, *Plos Genet*, submitted.
S21. T. Wicker *et al.*, *Nat Rev Genet* **8**, 973-982 (2007).
S22. B. Morgenstern, *Nucleic Acids Res* **32**, W33-36 (2004).
S23. C. Liang, L. Mao, D. Ware, L. Stein, *Genome Res* **19**, 1912-1923 (2009).
S24. A. A. Salamov, V. V. Solovyev, *Genome Res* **10**, 516-522 (2000).
S25. M. Spannagl, G. Haberer, R. Ernst, H. Schoof, K. F. Mayer, *Methods Mol Biol* **406**, 137-159 (2007).
S26. Y. Yu, C. Soderlund, *Plos Genet*, submitted.
S27. N. N. Alexandrov *et al.*, *Plant Mol Biol* **69**, 179-194 (2009).
S28. S. Ouyang *et al.*, *Nucleic Acids Res* **35**, D883-887 (2007).
S29. A. H. Paterson *et al.*, *Nature* **457**, 551-556 (2009).
S30. A. J. Vilella *et al.*, *Genome Res* **19**, 327-335 (2009).
S31. D. Swarbreck *et al.*, *Nucleic Acids Res* **36**, D1009-1014 (2008).
S32. S. Wessler, J. Bennetzen, R. K. Dawe, P. SanMiguel, N. Jiang, (National Science Foundation, 2006).
S33. J. L. Bennetzen, H. Wang, L. Yang, Q. Zhu, *Nature* **106**, 12832 (2009).
S34. R. Liu *et al.*, *Proc Natl Acad Sci U S A* **104**, 11844-11849 (2007).
S35. J. L. Bennetzen, C. Coleman, R. Liu, J. Ma, W. Ramakrishna, *Curr Opin Plant Biol* **7**, 732-736 (2004).
S36. S. C. Potter *et al.*, *Genome Res* **14**, 934-941 (2004).
S37. N. J. Mulder *et al.*, *Nucleic Acids Res* **31**, 315-318 (2003).
S38. A. Krogh, B. Larsson, G. von Heijne, E. L. Sonnhammer, *J Mol Biol* **305**, 567-580 (2001).
S39. H. Nielsen, J. Engelbrecht, S. Brunak, G. von Heijne, *Int J Neural Syst* **8**, 581-599 (1997).
S40. G. S. Slater, E. Birney, *BMC Bioinformatics* **6**, 31 (2005).

S41. E. Coe *et al.*, *Plant Physiol* **128**, 9-12 (2002).
S42. A. Sharma, G. G. Presting, *Mol Genet Genomics* **279**, 133-147 (2008).
S43. A. Sharma, K. L. Schneider, G. G. Presting, *Proc Natl Acad Sci USA* **105**, 15470-15474 (2008).
S44. T. K. Wolfgruber *et al.*, *Plos Genet*, submitted.
S45. K. M. Hufford, P. Canaran, D. H. Ware, M. D. McMullen, B. S. Gaut, *Plant Physiol* **144**, 1642-1653 (2007).
S46. B. J. Haas, A. L. Delcher, J. R. Wortman, S. L. Salzberg, *Bioinformatics* **20**, 3643-3646 (2004).
S47. A. H. Paterson, J. E. Bowers, B. A. Chapman, *Proc Natl Acad Sci USA* **101**, 9903-9908 (2004).
S48. X. Wang, H. Tang, J. E. Bowers, F. A. Feltus, A. H. Paterson, *Genetics* **177**, 1753-1763 (2007).
S49. M. Ashburner *et al.*, *Nat Genet* **25**, 25-29 (2000).
S50. Y. Benjamini, Y. Hochberg, *J Roy Stat Soc B Met* **57**, 289-300 (1995).
S51. S. Maere, K. Heymans, M. Kuiper, *Bioinformatics* **21**, 3448-3449 (2005).
S52. L. E. Palmer *et al.*, *Science* **302**, 2115-2117 (2003).
S53. J. A. Bedell *et al.*, *PLoS Biol* **3**, e13 (2005).
S54. J. H. Shi, R. K. Dawe, *Genetics* **173**, 1571-1583 (2006).
S55. N. G. Taylor, R. M. Howells, A. K. Huttly, K. Vickers, S. R. Turner, *Proc Natl Acad Sci U S A* **100**, 1450-1455 (2003).
S56. K. S. Dhugga *et al.*, *Science* **303**, 363-366 (2004).
S57. A. H. Liepman, C. G. Wilkerson, K. Keegstra, *Proc Natl Acad Sci U S A* **102**, 2221-2226 (2005).
S58. J. C. Cocuron *et al.*, *Proc Natl Acad Sci U S A* **104**, 8550-8555 (2007).
S59. B. Favery *et al.*, *Genes Dev* **15**, 79-89 (2001).
S60. C. M. Kim *et al.*, *Plant Physiol* **143**, 1220-1230 (2007).
S61. R. A. Burton *et al.*, *Science* **311**, 1940-1942 (2006).
S62. B. Penning *et al.*, *Plant Physiol*, in press.
S63. S. J. Emrich *et al.*, *Genetics* **175**, 429-439 (2007).
S64. N. M. Springer *et al.*, *Plos Genet*, in press.

Supporting Tables

Table S1. Estimation of maximum rate of sequencing errors. The maximum rate of sequencing errors was estimated with as a gold standard 10 genes (start to stop) cloned from B73 and sequenced by the Schnable Lab. The sequences of each gene had been carefully and manually edited and subsequently aligned to MF and HC reads (*S1*) during which process a few remaining errors were corrected. Eight of the ten genes aligned almost perfectly to the B73 reference genome. These genes, comprising 32 kb of aligned sequence were used to estimate the maximum rate of sequencing errors.

Name	**Accession number**	**Positions on B73 reference genome**	**% Coverage**	**% Identity**	**Length (start/stop) in bp**	**No. mismatches (bp)**	**Max. Error rate**
pdc2	AF370004	chr8_15591201-15594279	100.00	100.00	3,079	0	0
rth3	AY265855	chr1_47632393-47634396	100.00	100.00	2,004	0	0
rth1	AY265854	chr1_252977273-252989602	96.89	99.98	12,726	3	0.024%
pdc3	AF370006	chr1_45675008-45677591	100.00	99.92	2,582	3	0.116%
rf2e1	AY374447	chr5_189566690-189570467	100.00	100.00	3,778	0	0
rf2c	AF348412	chr3_220111752-220115384	100.00	99.97	3,632	2	0.055%
rf2d	AF348414	chr3_220129629-220131604	100.00	100.00	1,976	0	0
rf2b	AF348418	chr4_165057735-165060342	100.00	100.00	2,608	0	0
					32,385	**8**	**0.025%**

Table S2. Transposable elements (TEs) in the B73 RefGen_v1

Super-family		No. of families[a]	No. of TEs (x1000)[b]	Coverage (Mb)[b]	No. of intact TEs with captured gene fragments	Fraction of genome (%)[b]
Class I	LTR/*Copia*	109	404	484	36	23.7
	LTR/*Gypsy*	134	477	948	168	46.4
	LTR/Unknown	163	222	92.9	221	4.5
	LINE	31	35	20	n.d.	1.0
	SINE	4	1.99	0.5	n.d.	0.0
	Total class I	441	1,140	1,546	425	75.6
Class II	CACTA	156	12.4	64.7	155	3.2
	hAT	230	31.8	23.4	23	1.1
	MLE[c]/*Stowaway*	127	14.0	2.3	2	0.1
	MULE	155	12.9	20.2	262	1.0
	PIF/Tourist	179	49.7	19.8	20	1.0
	Helitron	8	22	45.5	1,194	2.2
	Total class II	855	143	176	1,656	8.6
Total TEs		1,296	1,283	1,722	2,081	84.2

[a] See supplemental text for family definitions.

[b] Intact plus fragmented, see supplemental text for definitions.

[c] *Mariner*-like element

Table S3. TE-related InterPro IDs used for exclusion of proteins from the Filtered Gene Set. Protein sequences were annotated for InterPro domains as described above. InterPro domains were chosen on the basis of association with independently-classified TE's and screened to ensure specificity.

InterPro ID	Domain Name
IPR018289	MULE transposase, conserved domain
IPR005162	Retrotransposon gag protein
IPR004242	Transposon, En/Spm-like
IPR002559	Transposase, IS4-like
IPR000477	RNA-directed DNA polymerase (reverse transcriptase)
IPR001584	Integrase, catalytic core
IPR013103	Reverse transcriptase, RNA-dependent DNA polymerase
IPR004332	Transposase, MuDR, plant
IPR013242	Retroviral aspartyl protease
IPR004252	Transposase, Ptta/En/Spm, plant
IPR009227	Zea mays MURB-like

Table S4. Minimum, maximum and average base-pair % coverage for each panel shown in Fig. S1. Values represent base-pair percent-coverage calculated in 1 Mb sliding windows, incrementing every 200 kb.

Chr	**Exon**			**LTR-RT**			**LTR-RT/gypsy**			**LTR-RT/copia**			**Non-CACTA**			**CACTA**		
	Min	**Max**	**Avg**	**Min**	**Max**	**Avg**	**Min**	**Max**	**Avg**	**Min**	**Max**	**Avg**	**Min**	**Max**	**Avg**	**Min**	**Max**	**Avg**
1	0	8.2	2.7	51.9	92.75	73.2	15.5	74.9	38.5	5.1	43.6	27.2	0.1	7.5	2.4	0	12.4	2.9
2	0	9.2	2.7	46.5	90.1	73.2	14	76	40.5	5.9	45.6	24.8	0.1	7.1	2.4	0.2	14.2	3.1
3	0	7.6	2.4	54.8	90.4	74.4	19.1	76.2	41.4	6.3	48.1	25.2	0	9.1	2.2	0.2	9	2.9
4	0	9.6	2.2	42.5	91.7	75.3	14.9	79.7	42.5	6.3	41.7	23.9	0	6.4	2.1	0.1	9.3	3.1
5	0	12.5	2.7	41	89.5	73.8	11.9	70.4	39.7	4.6	44.1	26.5	0	8.1	2.2	0.3	12.2	2.9
6	0	9	2.6	47.5	92.7	74.1	12.5	67	39.5	3.8	40.3	25.3	0	6.9	2.2	0.1	12.8	2.9
7	0	9.3	2.4	49.1	90.3	74.4	12.6	78.6	41.8	6.5	42.4	24.7	0	6.6	2.2	0.3	7.9	3.1
8	0.1	9.8	2.5	42.1	90.1	73.8	13.2	70.7	39.8	6.8	41.5	26.1	0.1	7.6	2.4	0.2	10.2	2.9
9	0.1	9.9	2.6	47.1	89.6	73.5	18.1	76.2	40.6	2.7	44.1	25	0	5.9	2.2	0.3	8.7	3
10	0.1	9	2.4	50.8	92.8	74.5	11.6	69.1	41.5	8.2	38.2	25	0.1	7	2.2	0.1	10.7	3.2

Table S5. Comparison of orthologous genes in maize, rice, sorghum, and Arabidopsis. Gene orthologs were defined phylogenetically with EnsemblCompara GeneTree method as described (*S30*). Sources of annotation: the maize filtered set; the sorghum genome (*S29*) JGI release Sbi 1.4 from March 2008, The Arabidopsis Information Resource (*S31*) release 8 from April 2008, and the MSU/TIGR Rice Genome Annotation Resource (*S28*) release 5 from January 2007.

	Maize[1]	**Sorghum**[2]	**Rice**[2]	**Arabidopsis**[2]
Gene num	27,741	20,408	20,498	18,434
Transcript num	48,170	22,038	28,747	23,118
Gene length (avg/median)	3982/2894	3458/2762	3685/3159	2526/2187
Transcript length	1684/1545	1562/1434	1872/1708	1667/1504
Protein length	358/303	430/375	430/373	442/382
Exon num	5.6/4	5.6/4	6.2/5	6.5/5
Exon length	302/156	280/147	300/151	255/144
Intron length	513/143	430/148	393/156	161/99
Gene GC	0.472	0.458	0.446	0.392
Exon GC	0.527	0.538	0.513	0.425
Intron GC	0.421	0.391	0.371	0.325
Genes with A/S	10442 (37.6%)	1320 (6.5%)	5125 (25%)	3563 (19.3%)
Max A/S	15	7	14	10
Single-Exon g	5407 (19.5%)	3879 (19%)	3608 (17.6%)	2606 (14.1%)
Max Exon num	53	58	78	79
Partial trans	2059	6495	0	2370

[1] Maize genes orthologous to rice, sorghum or *Arabidopsis*

[2] Genes orthologous to maize

Table S6. Summary of introns with repeats. The RefGen_v1 sequence was masked with the TE Consortium repeat library. A subset of genes (21,491) built purely from FLcDNAs were examined. Introns and genes that contain repeats were counted and percentages given.

	Intron count (gene count)	**Count of introns (genes) bearing repeats**	**Percentage of introns (genes) bearing repeats**
Total	80,561 (21,491)	11,265 (7,150)	14% (33.3%)
Intron length ≥1000 bp	10,776 (74,90)	7,170 (5,183)[a]	66.5% (69.2%)
Intron length <1000 bp	69,785 (14001)	4,095 (4,095)	5.9% (29.2%)

[a] Repeats ≥ 1000 bp: in 2399 introns (1,878 genes)

Table S7. Comparison of annotations generated by exclusive use FLcDNA versus the maize filtered set and other annotation sets.

	Maize cDNA[1]	**Maize All[2]**	**Rice All[3]**	**Maize cDNA ctg182[4]**
Gene num	20,867	32540	41,908	148
Transcript num	27,764	53,764	52,177	157
Gene length (avg/median)	3,521/2,495	3,757/2658	2,778/2,086	3,453/2,377
Transcript length	1,433/1,397	1,627/1492	1,502/1,355	1,542/1,515
Protein length	287/255	343/286	355/288	369/345
Exon num	4.7/4	5.3/4	4.9/3	4.7/4
Exon length	303/165	304/156	307/156	331/166
Intron length	581/148	516/146	412/170	499/154
Gene GC	0.471	0.471	0.452	0.481
Exon GC	0.534	0.527	0.516	0.557
Intron GC	0.425	0.421	0.376	0.421
Genes with A/S	4,971 (23.8%)	10,896 (33.5%)	6,387 (15.2%)	9 (6.1%)
Max A/S	10	15	14	2
Single-Exon g	4,877 (23.4%)	7,357 (22.6%)	9,710 (23.2%)	31 (20.9%)
Max Exon num	32	53	78	17
Partial trans	2,153	2,181	4	9

[1] Maize genes based on only FLcDNAs

[2] All maize genes in the filtered set

[3] All rice non-TE related genes annotated by TIGR Release 5.

[4] Selected maize genes based on FLcDNA predicted on ctg182

*~20% of the genes can be extended by species-specific or cross-species ESTs at 5’-end, which leads to longer CDS, suggesting those FLcDNAs might not be full-length.

Table S8. Quality assessment of 16 predicted starch-pathway genes in the filtered set. Established gene-structures from the maize starch biosynthetic pathway were chosen to assess the quality of the annotations in the maize filtered set.

Predicted CDS correctness	**Prediction on BAC contigs**	**Projected to AGP**	**Alignment on BAC**	**Alignment on AGP**	**Reason of incorrect prediction in filtered set**
Perfect match	9	8	complete	complete	Mis-projection
Near perfect	1	1	near complete	near complete	cDNA aligned to genome with low sequence identity
Lack of N-term	2	2	incomplete	incomplete	incomplete genome sequence
Extended N-term	1	1	complete	complete	incorrect start codon prediction
Lack of C-term	1	2	incomplete	incomplete	incomplete genome sequence
Split prediction[1]	2	2	incomplete	complete	unassembled contigs
w/ residue insertion/deletion	1	1	incomplete	incomplete	*ab initio* (non-evidence-based) prediction

[1] Two split genes were predicted for one authentic gene. For this reason the total number of genes listed as predictions is one larger than the number of authentic genes analyzed.

Table S9. Validation of gene structures of filtered gene set (FGS). Coding sequences of 10 gold standard genes were aligned to the B73 reference genome at ≥96% identity and 100% coverage. All filtered genes found within the same intra-BAC contig were projected onto the original BAC sequence from which the contig was derived. Seven of the ten genes have complete coverage in the filtered gene set. Only a partial gene models overlaps with the last exon of *rf2d* and no gene model was found for *pdc3*. Both genes have EST and FLcDNA support, do not contain known repeats within the annotated gene models and are located in the middle of the intra-BAC contigs. Two genes (4 coding sequences) map to *rf2a*, however, intron 5 was skipped in the model. The ~7kb intron 5 of *rf2a* contains retrotransposon sequences. It is conceivable that the associated gene models were truncated for reasons of size and/or repeat content. However FLcDNAs for this gene are available in Genbank as indicated in the body of the table.

Gene	**GenBank Accession**	**Filtered Gene Set Accession**	**Filtered Gene Set CDS**	**No. Inconsistencies**	**Comments**
gl8a	AF302098[a] U89509[b]	GRMZM2G087323	GRMZM2G087323_T01	1	FGS includes 4 extra bases after exon 1
pdc2	AF370004[a] AF370003[b]	GRMZM2G038821	GRMZM2G038821_T01	1	FGS includes extra 141bp at the beginning of exon 1
pdc3	AF370006[a] AF370005[b]	GRMZM2G385021	GRMZM2G385021_T01	8	FGS CDS does not align to gene sequence. FGS missing coverage of all 7 exons
rf2a	AF215823[a] U43082[b]	GRMZM2G058675 GRMZM2G072755	GRMZM2G058675_T01 GRMZM2G058675_T02 GRMZM2G072755_T01 GRMZM2G072755_T02	4	FGS missing 6th exon; 3 regions exhibit inconsistencies
rf2b	AF348418[a] AF348417[b]	GRMZM2G125268	GRMZM2G125268_T01 GRMZM2G125268_T02 GRMZM2G125268_T03	3	2 FGS CDS contain extra exonic sequences. 1 base in prediction inconsistent at beginning of 8th exon
rf2c	AF348412[a] AF348413[b]	GRMZM2G071021	GRMZM2G071021_T01 GRMZM2G071021_T02 GRMZM2G071021_T03	0	1 FGS CDS covers entire gene; 2 remaining CDS partially cover gene
rf2d	AF348414[a] AF348415[b]	GRMZM2G097699	GRMZM2G097699_T01	8	FGS contains extra 171bp at the beginning of exon 1 and 107bp at the end of exon 1; misses exons 2-7
rf2e	AY374447[a,c]	GRMZM2G169458	GRMZM2G169458_T01 GRMZM2G169458_T02	0	2 FGS CDS cover gene. Longer CDS accurately covers entire gene; shorter CDS has extra 91 bases predicted as exonic.
rth1	AY265854[a,c]	GRMZM2G099056	GRMZM2G099056_T01	0	FGS CDS accurately covers entire gene

Gene	GenBank Accession	Filtered Gene Set Accession	Filtered Gene Set CDS	No. Inconsistencies	Comments
rth3	AY265855[a,c]	GRMZM2G377215	GRMZM2G377215_T01 GRMZM2G377215_T02	0	2 FGS CDS cover gene. Longer CDS accurately covers entire gene; shorter CDS is missing 92bp in the middle of exon.

[a] Genomic sequence

[b] cDNA sequence

[c] CDS extracted from annotation of genomic sequences for this accession

Table S10. Correspondence between coding sequences (CDS) known from full-length cDNA clones and computational predictions from the maize genome sequence. Enzymes of the starch biosynthesis pathway were analyzed. cDNAs are identified by Genbank accession numbers (gi) and by genetically defined loci in those instances where mutations have been identified. "First nt." and "Last nt." indicate the starting and ending positions of the predicted gene in the chromosome pseudomolecule. Numbers in the correspondence description refer to the known coding sequence inferred from the cDNA.

	Known CDS			**Predicted CDS from genome sequence**					
Enzyme	**Locus**	**cDNA (gi)**	**Codons**	**Gene ID**	**Chr**	**First nt.**	**Last nt.**	**Strand**	**Correspondence between known and predicted proteins**
Starch synthase:									
GBSSI	*wx*	198442835	605	024993	9	23,213,761	23,213,761	1	Perfect
GBSSIb	n.d.	195647263	605	008263	7	37,074,030	37,081,289	-1	Perfect
SSI	n.d.	162458604	640	129451	9	17,537,935	17,547,474	1	Prediction lacks N terminal 168 residues; missing section not in gene set
SSIIa	*su2*	2811133	732	348551	6	113,380,769	113,385,825	1	Prediction has 60 extra N terminal residues; perfect from 1-176; ~20 residue insertion in predicted; perfect 177-257; mismatch from 258-275; perfect from 276-355; deletion of 356-410 in predicted; perfect from 410-729
SSIIb	n.d.	162463587	706	105791	5	205,769,894	205,774,702	-1	Near perfect, two residue deletion in prediction
SSIIc	n.d.	167860169	775	126988	5	32,364,622	32,366,243	-1	Prediction lacks N terminal 474 residues; missing section not in gene set
SSIII	*du1*	162463770	1674	141399	10	43,570,189	43,581,767	1	Perfect
SSIIIb	n.d.	145202747	1191	121612	2	142,647,341	142,656,379	-1	Perfect
SSIV	n.d.	194306594	909	044744	8	123,924,441	123,932,986	1	Perfect

	Known CDS			Predicted CDS from genome sequence					
Enzyme	**Locus**	**cDNA (gi)**	**Codons**	**Gene ID**	**Chr**	**First nt.**	**Last nt.**	**Strand**	**Correspondence between known and predicted proteins**
Starch branching enzyme:									
SBEI	*sbe1a*	162460641	823	088753	5	62,617,689	62,625,618	-1	Perfect
SBEIIa	*sbe2a*	TA176385 _4577*	844	073054	2	58,470,011	58,473,634	1	Prediction lacks C terminal 534 residues; missing section not in gene set
SBEIIb	*ae*	*162459705*	799	032628	5	167,879,647	167,690,099	-1	Prediction lacks C terminal 293 residues; missing section not in gene set
Starch debranching enzyme:									
ISA1	*su1*	195647079	789	138060	4	41,189,018	41,197,807	1	Perfect
ISA2	n.d.	162463512	729	090905	6	144,743,201	144,745,581	1	Perfect
ISA3	n.d.	162458750	694	150796	7	123,356,563	123,364,923	-1	Prediction has 83 extra N terminal residues
PU1	*zpu1*	3411265	962	353533 158043	2	104,979,675	104,986,054	-1	Prediction lacks C-terminal 400 residues; adjacent gene contains residues 608-755, then extends into predicted codons that do not match any known cDNA; missing sections not in gene set

*Full length mRNA sequence for this enzyme is not present in Genbank but is available as the indicated transcript assembly in the TIGR database.

Table S11. Top 50 InterPro hits in the Filtered Gene Set. Shown are counts and percents of genes bearing the listed domain.

ID	Domain Name	Count	Ratio
IPR011009	Protein kinase-like	1,298	2.32%
IPR002290	Serine/threonine protein kinase	1,155	2.06%
IPR001245	Tyrosine protein kinase	1,091	1.95%
IPR000104	Antifreeze protein, type I	910	1.63%
IPR008271	Serine/threonine protein kinase, active site	884	1.58%
IPR017442	Serine/threonine protein kinase-related	853	1.52%
IPR017441	Protein kinase, ATP binding site	804	1.44%
IPR001841	Zinc finger, RING-type	530	0.95%
IPR009057	Homeodomain-like	486	0.87%
IPR002885	Pentatricopeptide repeat	478	0.85%
IPR003593	ATPase, AAA+ type, core	378	0.68%
IPR016040	NAD(P)-binding	373	0.67%
IPR001611	Leucine-rich repeat	354	0.63%
IPR016024	Armadillo-type fold	352	0.63%
IPR001680	WD40 repeat	295	0.53%
IPR001005	SANT, DNA-binding	290	0.52%
IPR011046	WD40 repeat-like	281	0.50%
IPR011016	Zinc finger, RING-CH-type	280	0.50%
IPR014778	Myb, DNA-binding	278	0.50%
IPR016196	Major facilitator superfamily, general substrate transporter	276	0.49%
IPR000504	RNA recognition motif, RNP-1	273	0.49%
IPR003591	Leucine-rich repeat, typical subtype	270	0.48%
IPR001128	Cytochrome P450	264	0.47%
IPR012336	Thioredoxin-like fold	261	0.47%
IPR017853	Glycoside hydrolase, catalytic core	240	0.43%
IPR001810	Cyclin-like F-box	239	0.43%
IPR007087	Zinc finger, C2H2-type	231	0.41%
IPR016177	DNA-binding, integrase-type	229	0.41%
IPR002401	Cytochrome P450, E-class, group I	228	0.41%
IPR018247	EF-HAND 1	219	0.39%
IPR001471	Pathogenesis-related transcriptional factor and ERF, DNA-binding	213	0.38%
IPR017973	Cytochrome P450, C-terminal region	212	0.38%
IPR013210	Leucine-rich repeat, N-terminal	211	0.38%
IPR002403	Cytochrome P450, E-class, group IV	209	0.37%
IPR011598	Helix-loop-helix DNA-binding	207	0.37%
IPR017972	Cytochrome P450, conserved site	199	0.36%
IPR002048	Calcium-binding EF-hand	190	0.34%
IPR001092	Basic helix-loop-helix dimerisation region bHLH	189	0.34%
IPR016027	Nucleic acid-binding, OB-fold-like	186	0.33%
IPR015880	Zinc finger, C2H2-like	175	0.31%
IPR003072	Orphan nuclear receptor, NOR1 type	173	0.31%
IPR018248	EF hand	167	0.30%
IPR010255	Haem peroxidase	161	0.29%
IPR002016	Haem peroxidase, plant/fungal/bacterial	161	0.29%
IPR014001	DEAD-like helicase, N-terminal	154	0.28%
IPR000823	Plant peroxidase	148	0.26%

ID	Domain Name	Count	Ratio
IPR003441	No apical meristem (NAM) protein	145	0.26%
IPR001965	Zinc finger, PHD-type	141	0.25%
IPR002213	UDP-glucuronosyl/UDP-glucosyltransferase	139	0.25%
IPR009072	Histone-fold	137	0.24%

Table S12. Possible sources of artifacts that explain rice-sorghum specific families. Peptide queries from rice-sorghum specific families were aligned to the indicated databases with TBLASTN (Wu-blast, P-value cutoff 1e-20). In total there were 108 non-TE rice genes and 96 non-TE sorghum genes that aligned to maize cDNA, but not to the maize genome sequence; indicating genes not annotated due to missing genomic sequence.

	Total	**Aligned to Plant TEs**	**Aligned to Maize Genome (TE)**	**Aligned to Maize cDNA (TE)**	**On Maize Genome-cDNA non-TE**
Rice Genes	1,220	184	406 (66)	476 (43)	325
Sorghum Genes	1,013	176	454 (171)	481 (125)	260
Rice families		57	259 (44)	303 (29)	206
Sorghum families		56	274 (55)	326 (45)	203

Table S13. Representation of tandem repeats and retrotransposons in the B73 RefGen_v1 and their enrichment in the ChIP reads. The numbers of nucleotides attributable to each tandem repeat in the whole genome shotgun (WGS) data, the B73 reference genome, and ChIP reads were determined with competitive cross_match. Competitive WU-BLAST was used to quantify CRM and opie elements, as well as genes, in each data set. Note that the CRM elements are well represented in the reference genome. Sequence representation in B73 RefGen_v1, and enrichment in the ChIP fraction were calculated for each data set by dividing the fraction of nucleotides represented by a sequence in B73 RefGen_v1 or the ChIP sequence by the fraction of nucleotides represented by that sequence in the WGS data.

Database	**Whole Genome Shotgun**		**B73 RefGen_v1**			**CENH3-Immunoprecipitated Chromatin**		
	Nucleotides	**Fraction**	**Nucleotides**	**Fraction**	**Representation**	**Nucleotides**	**Fraction**	**Enrichment**
Tandem Repeats								
CentC NCBI 79	1,900,089	1.75E-03	1,947,833	9.51E-04	54.41%	1,447,866	5.85E-02	33.5
Cent4	50,263	4.62E-05	59,875	2.92E-05	63.23%	1,935	7.82E-05	1.69
Knob180	14,799,588	1.36E-02	2,520,776	1.23E-03	9.04%	419,023	1.69E-02	1.24
Knob350	1,047,788	9.64E-04	377,601	1.84E-04	19.13%	30,428	1.23E-03	1.28
rDNA 5S B73	64,953	5.98E-05	38,562	1.88E-05	31.51%	2,801	1.13E-04	1.90
rDNA 45S B73	17,749,102	1.63E-02	4,505,343	2.20E-03	13.47%	198,428	8.02E-03	0.49
Retrotransposons and Genes								
CRM1	2,951,440	2.72E-03	5,960,580	2.91E-03	107.19%	1,091,992	4.42E-02	16.3
CRM2	1,341,433	1.23E-03	2,674,521	1.31E-03	105.82%	1,560,396	6.31E-02	51.1
CRM3	287,701	2.65E-04	549,878	2.68E-04	101.44%	127,145	5.14E-03	19.4
CRM4	1,710,355	1.57E-03	3,603,311	1.76E-03	111.82%	46,815	1.89E-03	1.20
Opie	37,681,548	3.47E-02	95,565,035	4.67E-02	134.61%	1,065,855	4.31E-02	1.24
Genes	1,482,349	1.36E-03	2,507,971	1.22E-03	89.80%	31,368	1.27E-03	0.93

Table S14. Counts of genes showing synteny in comparisons of maize to sorghum and maize to rice, and their percentage relative to orthologous genes used as input for this analysis.

Genome (comparison)	**Gene Counts**		**% Syntenic**
	Syntenic	**Orthologous**	
Maize (total)	23,589	27,550	85.6
Maize (vs sorghum)	21,766	25,216	86.3
Maize (vs rice)	19,579	25,844	75.8
Sorghum (vs maize)	18,116	20,408	89.8
Rice (vs maize)	15,750	20,569	76.6

Table S15. Enrichment for functional classes in retained duplicate homoeologues. Enrichment for functional classes in retained duplicate homoeologs and in singletons was tested with the GO Molecular Function and GO Biological Process terms (*S49*). Those with significant P-values are shown.

GO Molecular Function

Retained homoeologues

GO ID	**GO description**	**Query Frequency**		**Total Frequency**		**p vlaue**
30528	transcription regulator activity	433/3,928	11.0%	1,231/17,210	7.1%	7.61E-22
3677	DNA binding	747/3,928	19.0%	2,402/17,210	13.9%	1.19E-21
3700	transcription factor activity	290/3,928	7.3%	806/17,210	4.6%	7.60E-16
3676	nucleic acid binding	999/3,928	25.4%	3,599/17,210	20.9%	6.72E-13
43565	sequence-specific DNA binding	207/3,928	5.2%	582/17,210	3.3%	1.36E-10
50824	water binding	323/3,928	8.2%	1,007/17,210	5.8%	2.51E-10
50825	ice binding	323/3,928	8.2%	1,007/17,210	5.8%	2.51E-10
5488	binding	2,807/3,928	71.4%	11,630/17,210	67.5%	1.26E-07
4871	signal transducer activity	222/3,928	5.6%	695/17,210	4.0%	8.25E-07
60089	molecular transducer activity	222/3,928	5.6%	695/17,210	4.0%	8.25E-07
8270	zinc ion binding	478/3,928	12.1%	1,688/17,210	9.8%	1.20E-06
3779	actin binding	40/3,928	1.0%	77/17,210	0.4%	1.47E-06
8092	cytoskeletal protein binding	40/3,928	1.0%	80/17,210	0.4%	5.38E-06
Singletons						
3824	catalytic activity	3,568/6,500	54.8%	8,846/17,210	51.4%	5.54E-10
16787	hydrolase activity	1,288/6,500	19.8%	2,978/17,210	17.3%	4.91E-09

GO Biological Process

Retained homoeologues

GO ID	**GO description**	**Query Frequency**		**Total Frequency**		**p vlaue**
65007	biological regulation	1,215/3,123	38.9%	3,812/13,439	28.3%	1.38E-45
60255	regulation of macromolecule metabolic process	706/3,123	22.6%	1,956/13,439	14.5%	7.09E-42
50791	regulation of biological process	923/3,123	29.5%	2,755/13,439	20.5%	1.67E-41
19222	regulation of metabolic process	713/3,123	22.8%	1,993/13,439	14.8%	4.45E-41
45449	regulation of transcription	671/3,123	21.4%	1,848/13,439	13.7%	7.61E-41
19219	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	676/3,123	21.6%	1,866/13,439	13.8%	7.61E-41
51244	regulation of cellular process	890/3,123	28.4%	2,662/13,439	19.8%	1.56E-39
10556	regulation of macromolecule biosynthetic process	671/3,123	21.4%	1,865/13,439	13.8%	1.72E-39

GO Biological Process

Retained homoeologues

GO ID	GO description	Query Frequency		Total Frequency		p vlaue
9889	regulation of biosynthetic process	671/3,123	21.4%	1,865/13,439	13.8%	1.72E-39
10468	regulation of gene expression	676/3,123	21.6%	1,884/13,439	14.0%	2.00E-39
31323	regulation of cellular metabolic process	680/3,123	21.7%	1,902/13,439	14.1%	3.96E-39
6355	regulation of transcription, DNA-dependent	419/3,123	13.4%	1,206/13,439	8.9%	8.90E-20
51252	regulation of RNA metabolic process	420/3,123	13.4%	1,210/13,439	9.0%	8.90E-20
9628	response to abiotic stimulus	339/3,123	10.8%	1,049/13,439	7.8%	9.38E-11
42309	homoiothermy	323/3,123	10.3%	1,007/13,439	7.4%	7.28E-10
9409	response to cold	323/3,123	10.3%	1,007/13,439	7.4%	7.28E-10
1659	temperature homeostasis	323/3,123	10.3%	1,007/13,439	7.4%	7.28E-10
50826	response to freezing	323/3,123	10.3%	1,007/13,439	7.4%	7.28E-10
9266	response to temperature stimulus	323/3,123	10.3%	1,009/13,439	7.5%	9.00E-10
32501	multicellular organismal process	345/3,123	11.0%	1,091/13,439	8.1%	9.07E-10
42592	homeostatic process	364/3,123	11.6%	1,174/13,439	8.7%	3.74E-09
65008	regulation of biological quality	374/3,123	11.9%	1,214/13,439	9.0%	4.68E-09

Table S16. Classification of sorghum loci with respect to syntenic positions in maize. Sorghum was used as a reference to assign syntenic maize genes to their ancestral homoeologous chromosomes, A and B. Each sorghum gene was then classified with respect to the disposition of corresponding sites in maize, either present as a singleton in A, singleton in B, or as duplicates retained in both A and B.

Chr.	**Sites**[a]	**A only (%)**[b]	**B only (%)**	**Both A & B**	**Fractionation Bias**[c]
Sb1	3503	58.4	22.5	19.1	2.60
Sb2	2378	59.3	25.4	15.3	2.33
Sb3	2753	56.8	25.9	17.3	2.19
Sb4	2213	54.9	27.8	17.3	1.97
Sb5	678	66.7	23.2	10.2	2.88
Sb6	1678	53.8	28.1	18.1	1.91
Sb7	1187	57.2	27.3	15.5	2.10
Sb8	846	58.8	28.3	12.9	2.08
Sb9	1529	58.3	27.3	14.4	2.14
Sb10	1540	62.4	25.0	12.6	2.50

[a] Corresponds to syntenic sorghum loci on the specified chromosome.

[b] "A" and "B" refer to ancestral homoeologous chromosomes in maize. "A only (%)" gives percent of sites that are located on A but not B; vice-versa for "B only (%)". The homoeologue with the lower fractionation rate was designated as "A".

[c] Ratio of "A only" to "B only"

Table S17. Raw data summarized in Table S16. This table shows syntenic gene counts broken out by maize chromosome components.

Sorghum Chr.	Syntenic Genes	Maize Ancestral Homoeolog	Maize Chr. Components	Singletons	Singletons (components summed)	Duplicates	Duplicates (components summed)	% singleton	% duplicate
Sb1	3503	A	Zm1	2,046	2,046	668	668	58.4	19.1
		B	Zm5	415	789	314	668	22.5	
			Zm9	374		354			
Sb2	2378	A	Zm7	1,410	1,410	365	365	59.3	15.3
		B	Zm2	603	603	365	365	25.4	
Sb3	2753	A	Zm3	1,564	1,564	477	477	56.8	17.3
		B	Zm8	712	712	477	477	25.9	
Sb4	2213	A	Zm5	1,215	1,215	383	383	54.9	17.3
		B	Zm4	615	615	383	383	27.8	
Sb5	678	A	Zm4	452	452	69	69	66.7	10.2
		B	Zm2	157	157	69	69	23.2	
Sb6	1685	A	Zm2	903	903	304	304	53.8	18.1
		B	Zm10	471	478	304	304	28.1	
			Zm3	7		0			
Sb7	1187	A	Zm1	409	679	116	184	57.2	15.5
			Zm10	179		36			
			Zm6	87		31			
			ZmUNKNOWN	4		1			
		B	Zm4	324	324	184	184	27.3	
Sb8	781	A	Zm1	177	457	50	104	58.8	12.9
			Zm10	264		47			
			Zm6	16		7			
		B	Zm3	220	220	104	104	28.3	
Sb9	1464	A	Zm10	56	854	13	211	58.3	14.4
			Zm6	798		198			
		B	Zm8	399	399	211	211	27.3	
Sb10	1474	A	Zm5	194	920	47	186	62.4	12.6
			Zm9	720		139			
			ZmUNKNOWN	6		0			
		B	Zm6	368	368	186	186	25.0	

Table S18. Top thirty paralog clusters and their distribution across homoeologous sister regions. Clusters were defined as proximally located paralogs that have no more than two intervening non-paralogous genes. Clusters were ranked on the basis of the sum of cluster sizes at both homoeologous sites.

Homoeologous Site A			**Homoeologous Site B**			**InterPro ID**	**InterPro description**
First gene	**chr**	**size**	**First gene**	**chr**	**size**		
GRMZM2G427301	4	20	-	-	-	IPR013865	Protein of unknown function DUF1754, eukaryotic
AC190772.4_FG011	4	19	-	-	-	IPR001929	Germin
GRMZM2G145069	1	10	GRMZM2G161827	9	3	IPR010987	Glutathione S-transferase, C-terminal-like
GRMZM2G042712	2	9	GRMZM2G012636	7	4	IPR003676	Auxin responsive SAUR protein
GRMZM2G160526	6	13	-	-	-	IPR011050	Pectin lyase fold/virulence factor
GRMZM2G418833	1	13	-	-	-	IPR009009	Barwin-related endoglucanase
GRMZM2G394500	10	10	GRMZM2G365774	8	2	IPR000823	Plant peroxidase
GRMZM2G427903	7	9	GRMZM2G044049	2	3	IPR000823	Plant peroxidase
GRMZM2G099737	7	11	-	-	-	IPR005174	Protein of unknown function DUF295
GRMZM2G130800	3	7	AC216871.3_FG001	8	3	IPR007657	Glycosyltransferase AER61, uncharacterised
GRMZM2G147752	3	6	GRMZM2G348090	8	4	IPR001128	Cytochrome P450
GRMZM2G094713	1	6	GRMZM2G091457	5	4	IPR001245	Tyrosine protein kinase
AC152495.1_FG002	10	10	-	-	-	IPR000767	Disease resistance protein
GRMZM2G079219	8	10	-	-	-	IPR001245	Tyrosine protein kinase
GRMZM2G094028	3	5	GRMZM2G171807	8	4	IPR001245	Tyrosine protein kinase
GRMZM2G013002	9	5	GRMZM2G021621	1	4	IPR005132	Rare lipoprotein A
GRMZM2G161306	1	9	-	-	-	IPR001052	Rubredoxin
GRMZM2G309258	7	6	GRMZM2G087625	2	2	IPR001245	Tyrosine protein kinase
GRMZM2G118809	1	5	GRMZM2G312069	5	3	IPR001128	Cytochrome P450
GRMZM2G146209	5	8	-	-	-	IPR000864	Proteinase inhibitor I13, potato inhibitor I
GRMZM2G152553	4	8	-	-	-	IPR013170	mRNA splicing factor, Cwf21
GRMZM2G088273	4	8	-	-	-	IPR002530	Zein seed storage protein
GRMZM2G069737	6	8	-	-	-	IPR001810	Cyclin-like F-box
AC234519.1_FG005	8	8	-	-	-	IPR018119	Strictosidine synthase, conserved region
GRMZM2G325023	5	8	-	-	-	IPR002213	UDP-glucuronosyl/UDP-glucosyltransferase
GRMZM2G470309	1	5	GRMZM2G178024	9	2	IPR004263	Exostosin-like
GRMZM2G416652	9	4	GRMZM2G376684	6	3	IPR009057	Homeodomain-like
GRMZM2G120794	7	4	GRMZM2G403590	2	3	IPR007087	Zinc finger, C2H2-type
GRMZM2G074604	5	4	GRMZM2G160541	4	3	IPR001106	Phenylalanine/histidine ammonia-lyase
GRMZM2G375602	3	6	GRMZM2G004947	8	1	IPR001087	Lipase, GDSL

Supporting Figures

Figure S1. Distribution of genes (exons) and transposable element repeat classes across the ZmB73v1 assembly. Base-pair percent-coverage was calculated in 1 Mb sliding windows, incrementing every 200 kb. Track “Non-CACTA DNA TE” includes Class II elements Mutator, hAT, Pif-Harbinger and Tc1-Mariner. Arrow indicates the position of centromeres. Scale (min = blue; max = red). See Table S4 for the range of minimum and maximum values for each track.

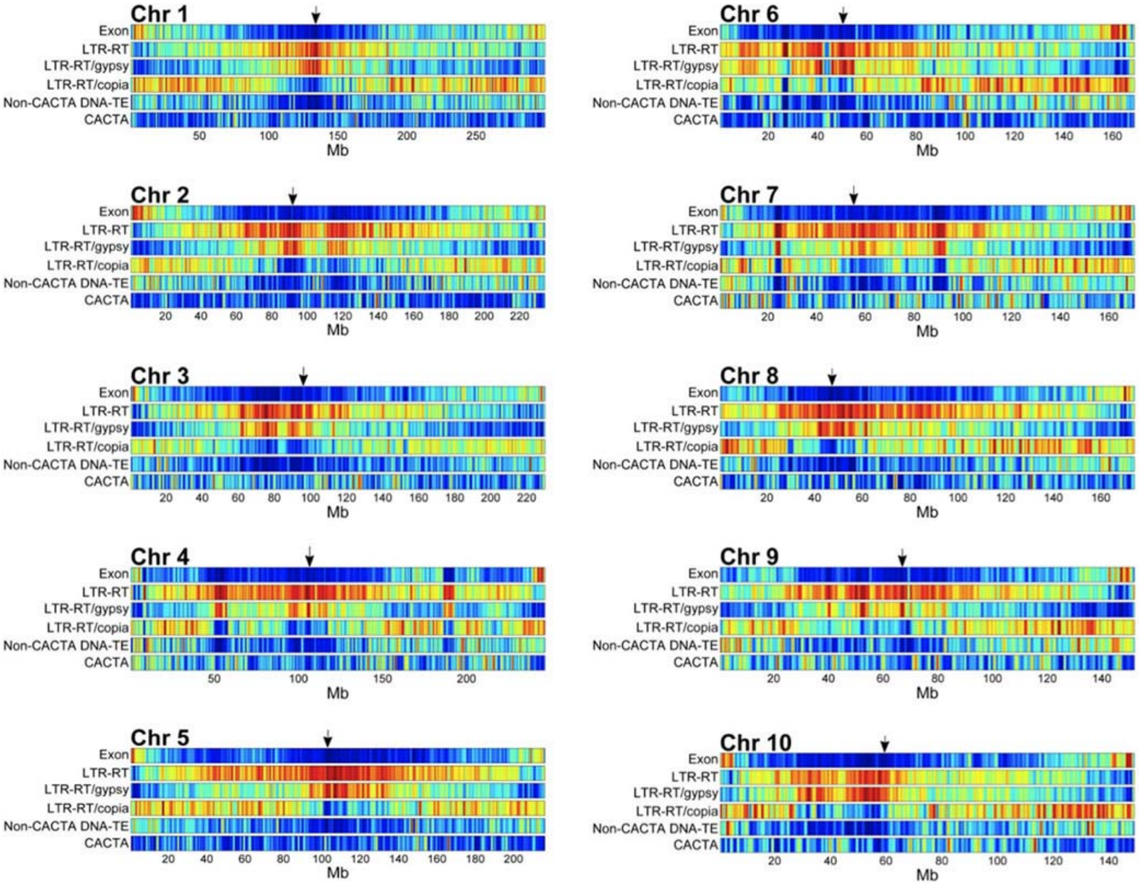


Figure S2. Process for selecting the filtered gene set on AGP. The projected genes are from direct coordinate mapping between intra-BAC contigs and the AGP. The mapped genes are those predicted on BAC contigs that are not used in the AGP; they are aligned to the AGP on the basis of sequence similarity. See supporting online text for details on the filtering process.

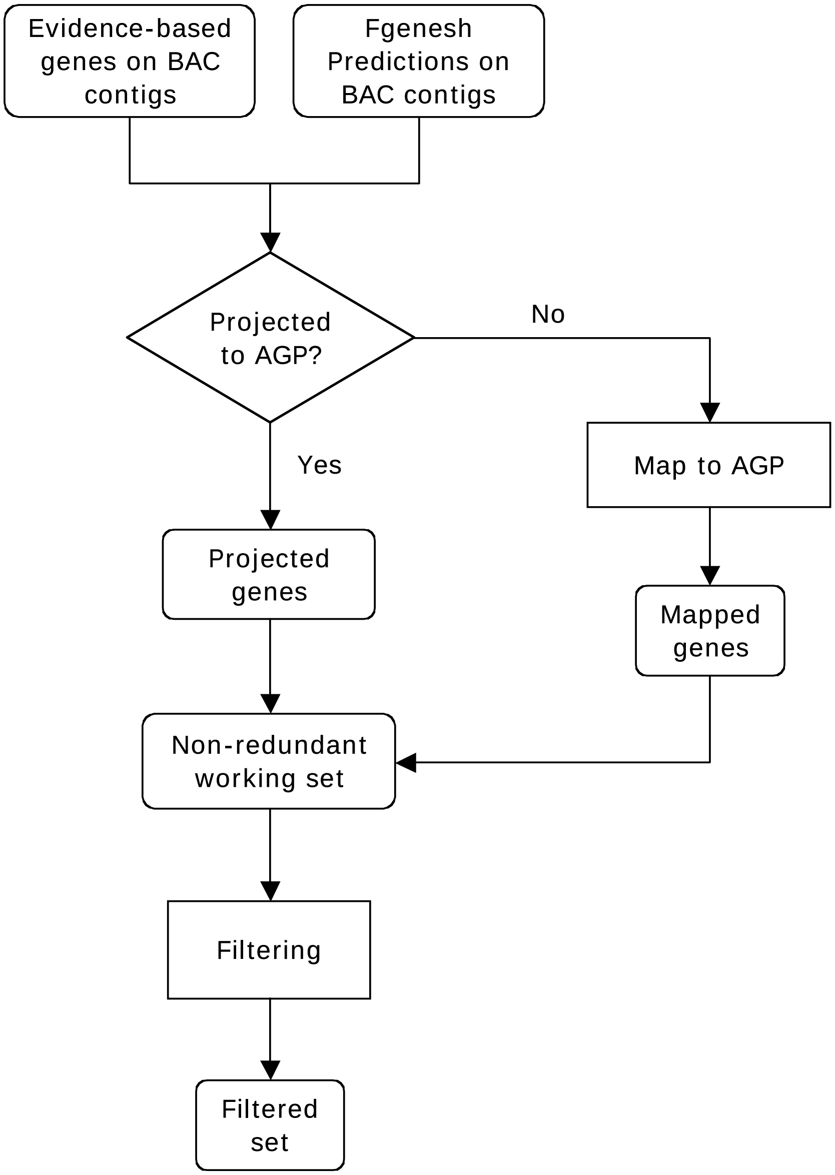

Evidence-based genes on BAC contigs
Fgenesh Predictions on BAC contigs
Projected to AGP?
No
Yes
Map to AGP
Projected genes
Mapped genes
Non-redundant working set
Filtering
Filtered set

Figure S3. Species comparison of exon length distribution. Arabi-zm-ort: indicates Arabidopsis genes with orthology to maize; rice-zm-ort indicates rice genes with orthology to maize; sorghum-zm-ort indicates sorghum genes with orthology to maize; maize-all indicates all maize genes in the filtered set; maize ort indicates maize gene with orthology to Arabidopsis, rice or sorghum genes; and maize-cdna: indicates maize genes built with pure fl-cDNAs.

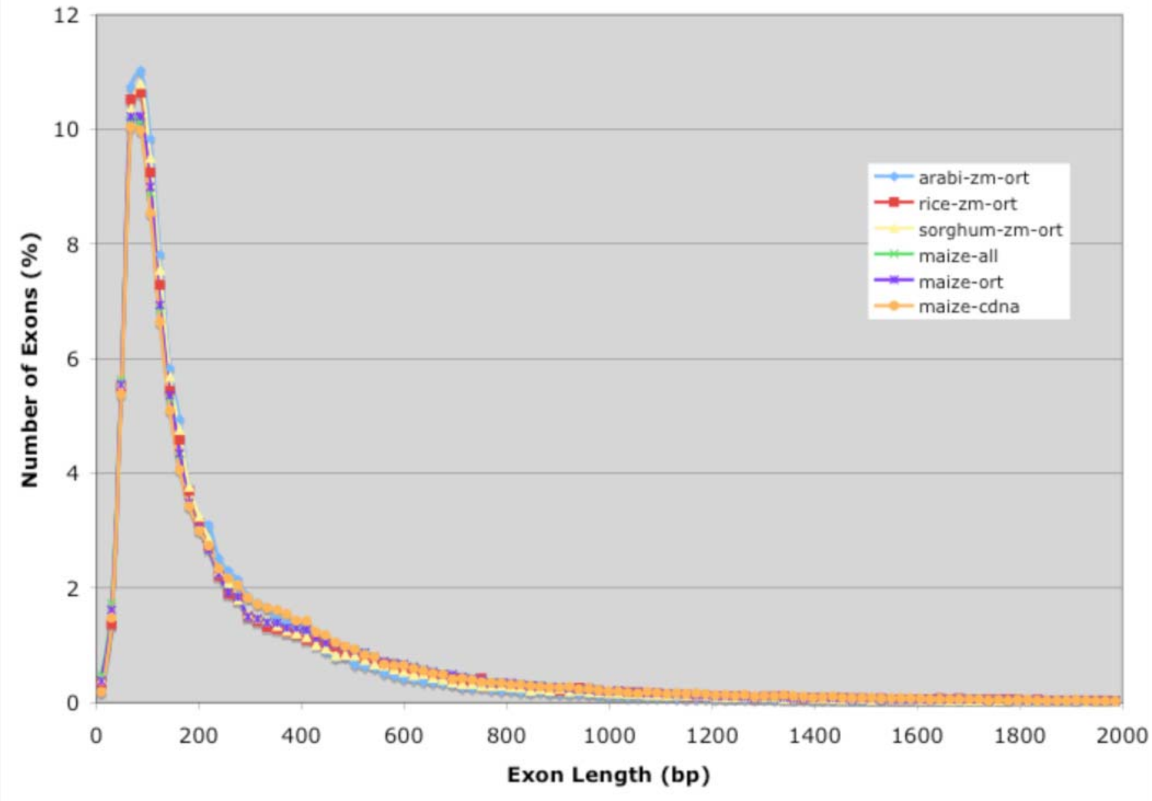


Figure S4. Species comparisons of intron length distributions. This cumulative distribution plot shows fraction of introns with lengths exceeding threshold. See Fig. S3 legend for explanation of gene sets used in this analysis.

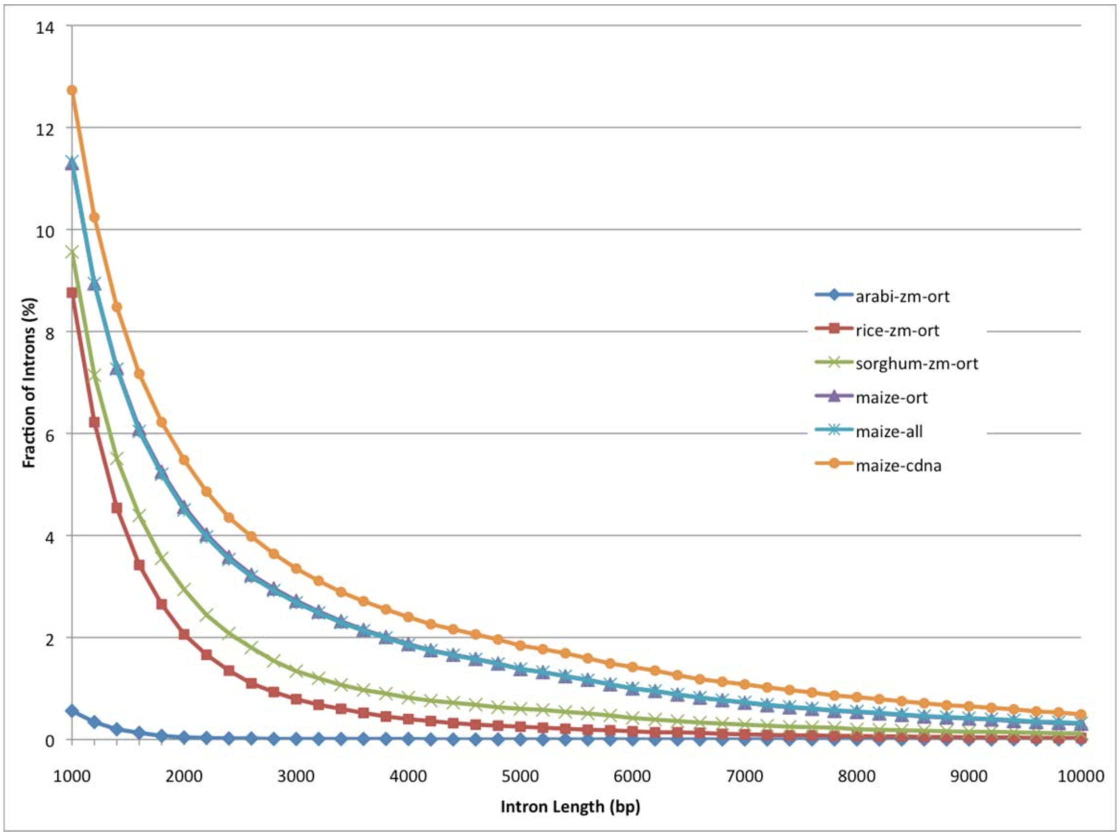


Figure S5. Exon GC percentage. See Fig. S3 legend for explanation of gene sets used in this analysis.

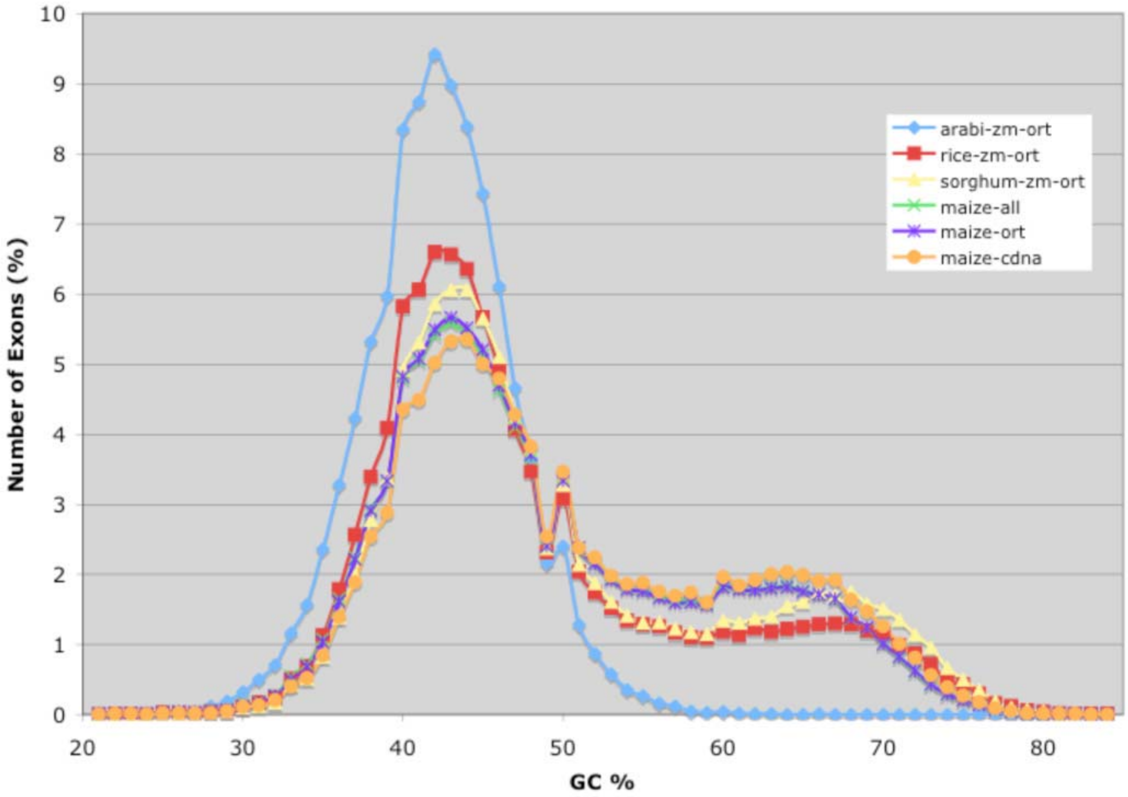


Figure S6. Intron GC Percentage. See Fig. S3 legend for explanation of gene sets used in this analysis.

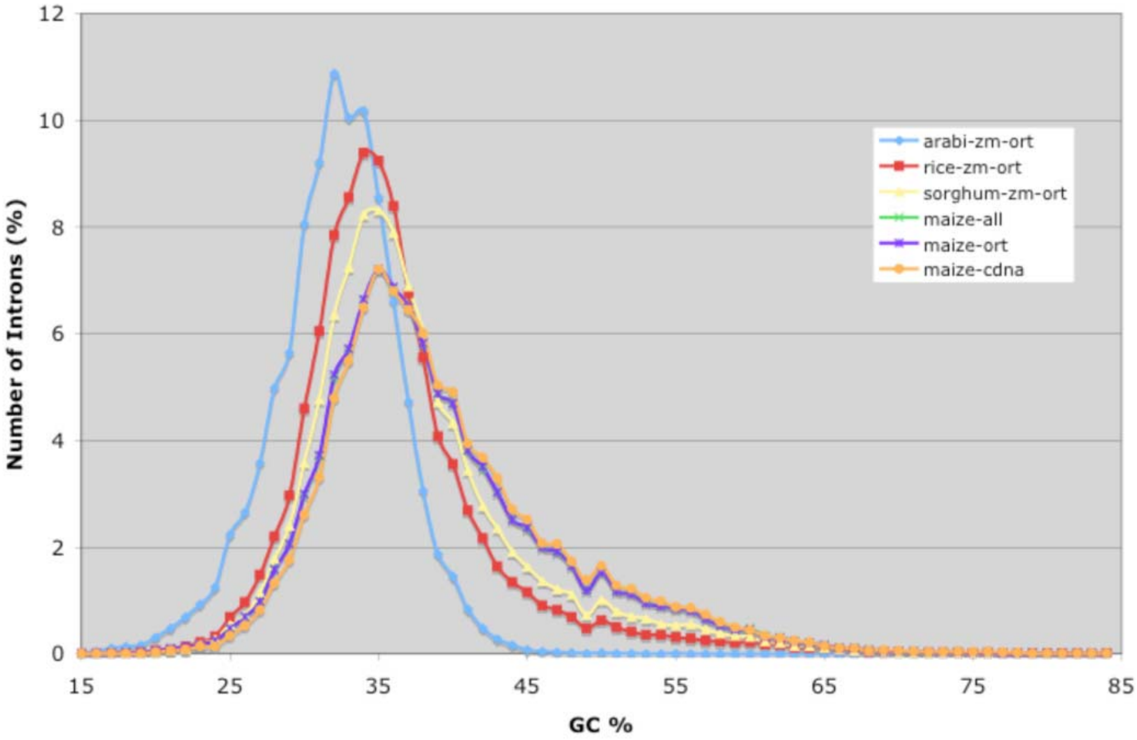


Figure S7. Transcript GC percentage. See Fig. S3 legend for explanation of gene sets used in this analysis.

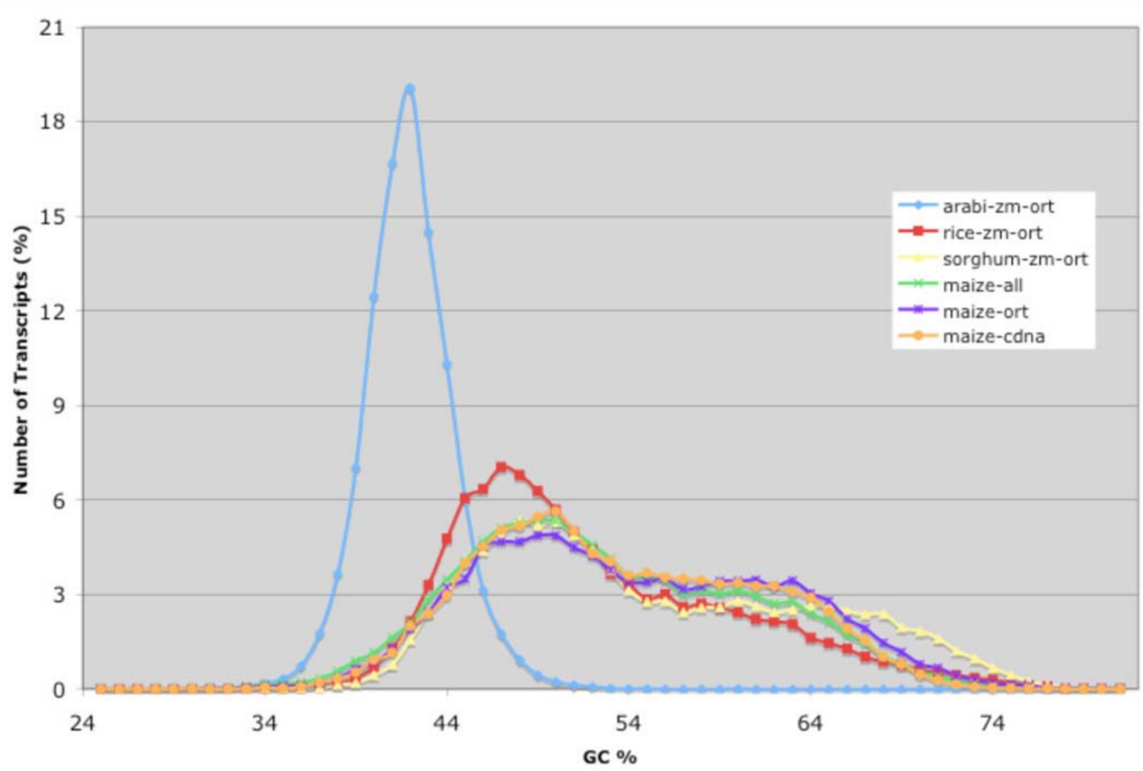


Figure S8. Comparison of protein length distributions among species annotations. See Fig. S3 legend for explanation of gene sets used in this analysis.

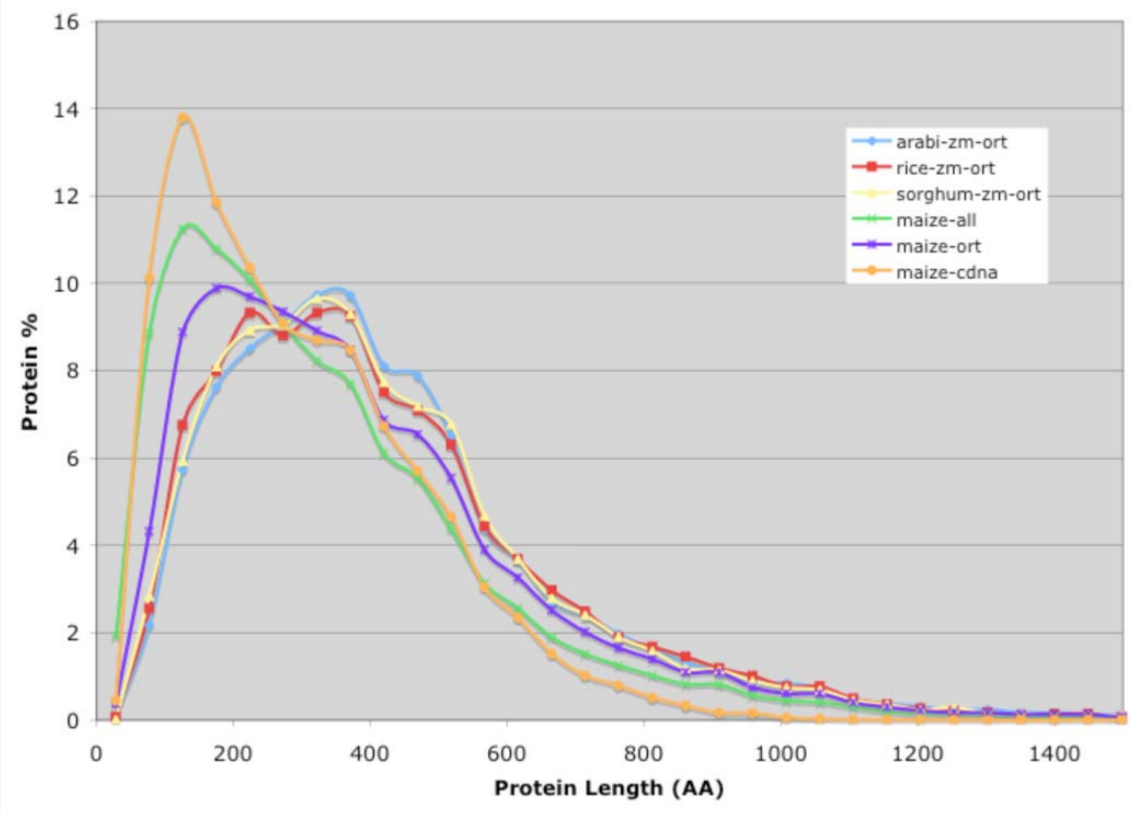


Figure S9. The 10 most common Gene Ontology (GO) terms in the Filtered Gene Set contains for each functional class.

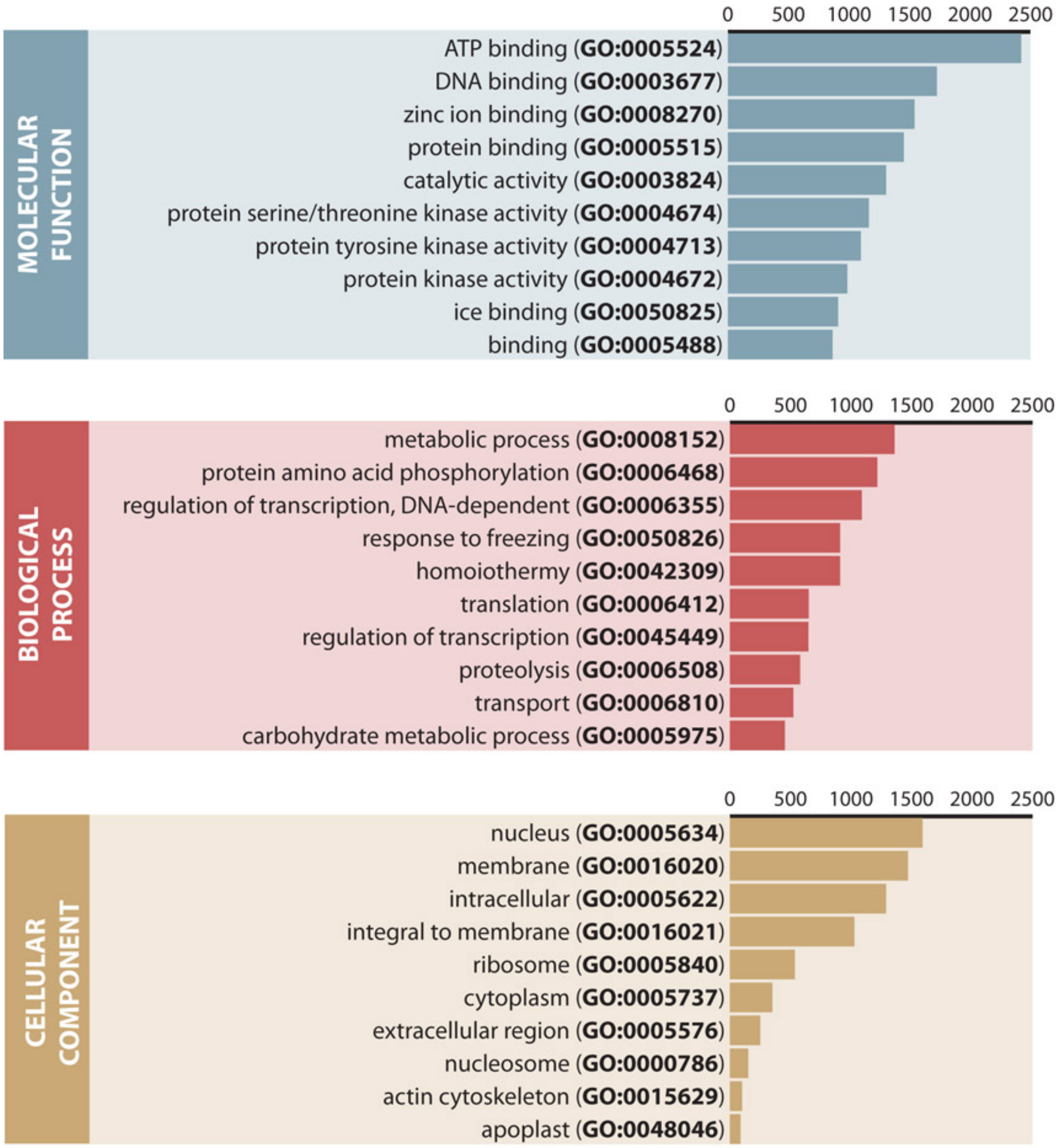


Figure S10. Distribution of RNA-seq reads from seedling leaf to the ZmB73v1. Approximate number of 32 nt reads from seedling leaves are shown (in millions) together with percentage of placed reads to each class. No placement data was obtained for approximately 22% of reads. Of the 32,540 genes in the filtered gene set, 26,857 evidence-based and 1,703 FGENESH models were verified. Percentages do not sum to 100 due to rounding.

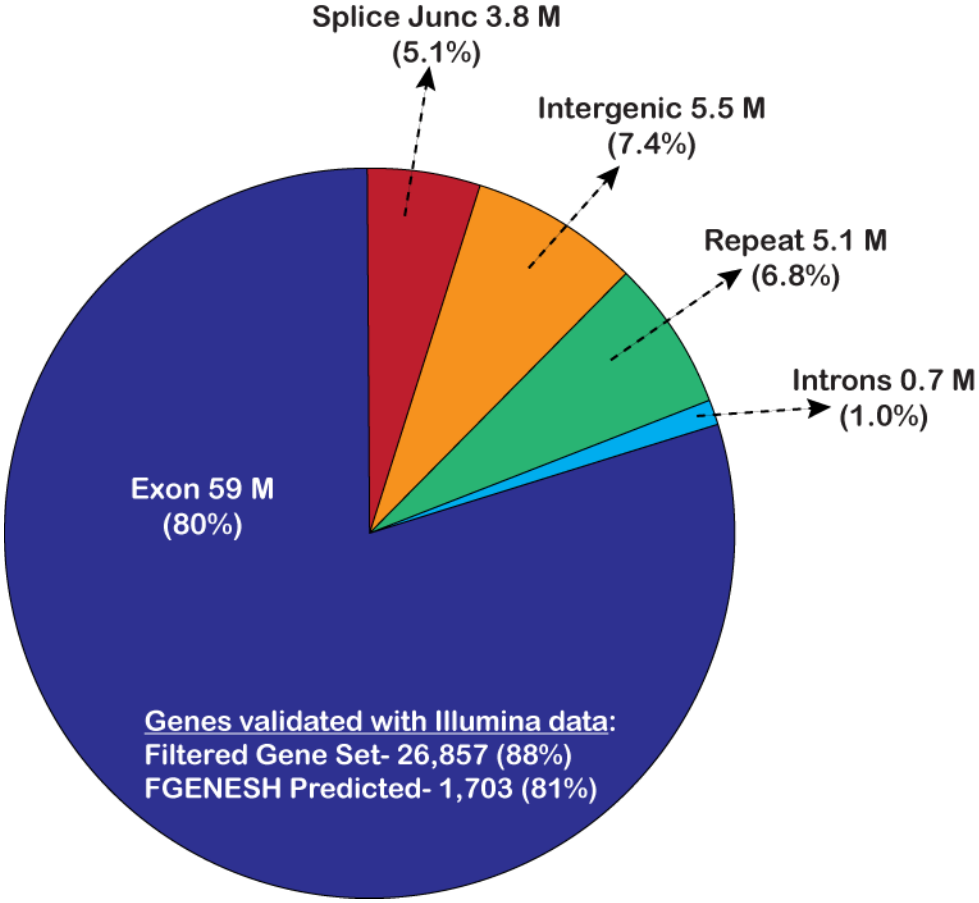


Figure S11. Number of genes from the filtered gene set (N= 32,540) whose transcription is supported by RNA-seq experiments. In combination the three RNA-seq experiments provide evidence for the transcription of 91% of the genes in the filtered gene set (29,541/32,540).

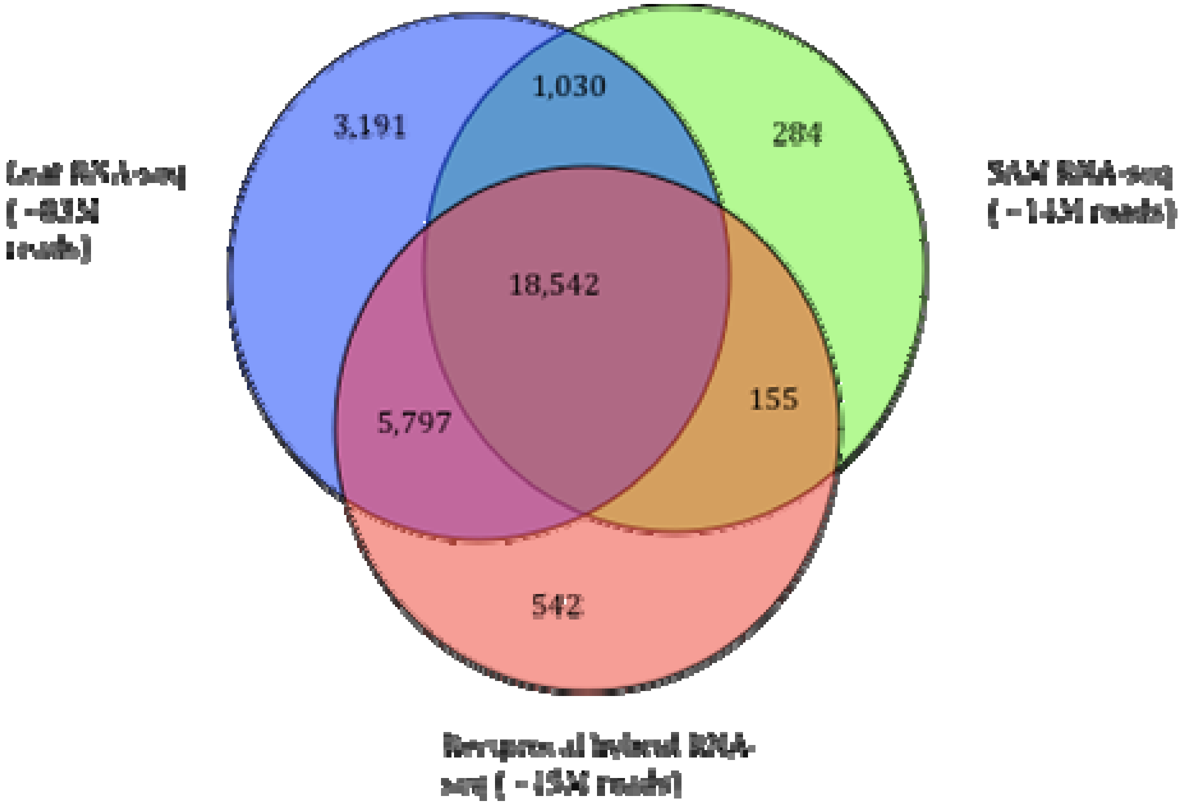


Figure S12. Centromere repeats of maize chromosomes. This composite image of chromosomes 1, 2, 3, 8 and 9 shows the position of the centromere on the basis of anti-CENH3 ChIP and the location of CRM and CentC repeats. CENH3 density is displayed as a moving average of the sum of the number of 454 reads per 0. 1% chromosome arm length for all 5 chromosomes, calculated over nine windows. Blue line = uniquely mapped reads at 100% identity, red line = single best match at ≥96% identity and ≥96% read length. CentC and CRM repeats are shown as the sum of the number of nucleotides per 0.1% chromosome arm length for all 5 chromosomes. Note the exponential y-axis for the CentC panel due to the high concentration of CentC in the centromeres and the variable scales for the CRM panels. Arrow = centromere.

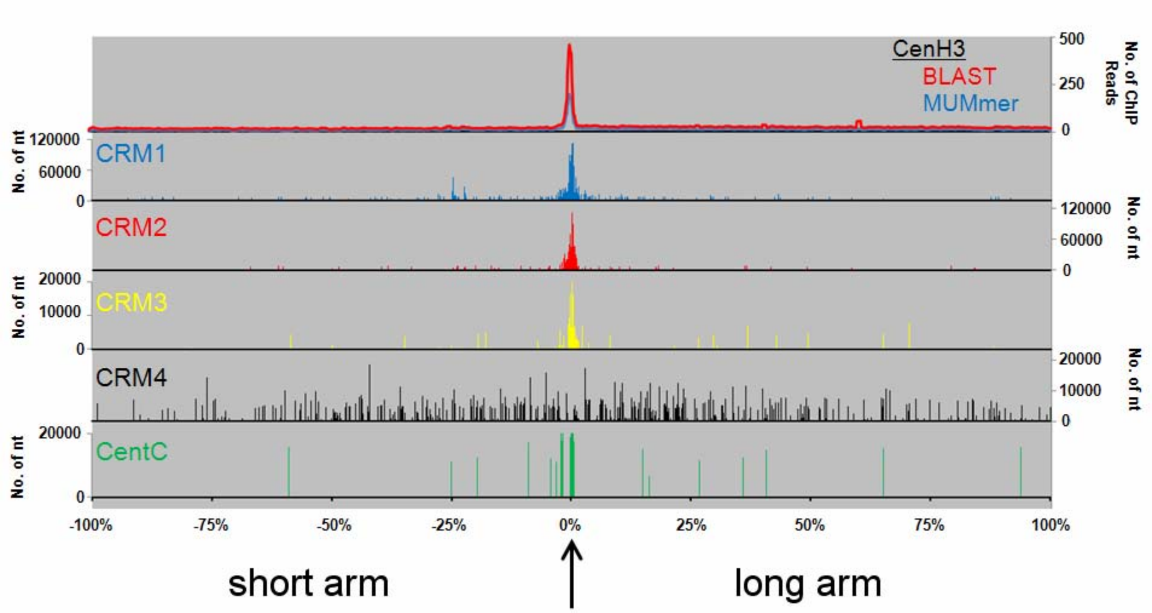


Figure S13. Distribution of DNA Methylation and Heterochromatin in Maize and Sorghum. A) Approximately 567,000 B73 methyl filtration reads were retrieved from Genbank (*S52*) and aligned to the ZmB73v1 assembly (CP) with BLAST. Chromosomes were divided into 0.1 Mb bins and the number of reads in each bin that aligned uniquely in the sequenced genome with at least 95% identity were counted. Per-bin counts were divided by the total number of reads mapping uniquely in the genome to yield relative enrichment or depletion scores for coverage in each bin. Red tones reflect enrichment in MF sequence density, while green tones reflect depletion. B) 534,000 sorghum bicolor reads (*S53*) were retrieved from Genbank and aligned to the SbiI sorghum genome (*S29*) with BLAST. Relative enrichment or depletion scores for coverage across 0.1 Mb bins was calculated as in maize. C) The MF density map for B73 chromosome 8 is compared to DAPI, CENPC, and H3K27me3 staining performed by Shi and Dawe (*S54*).

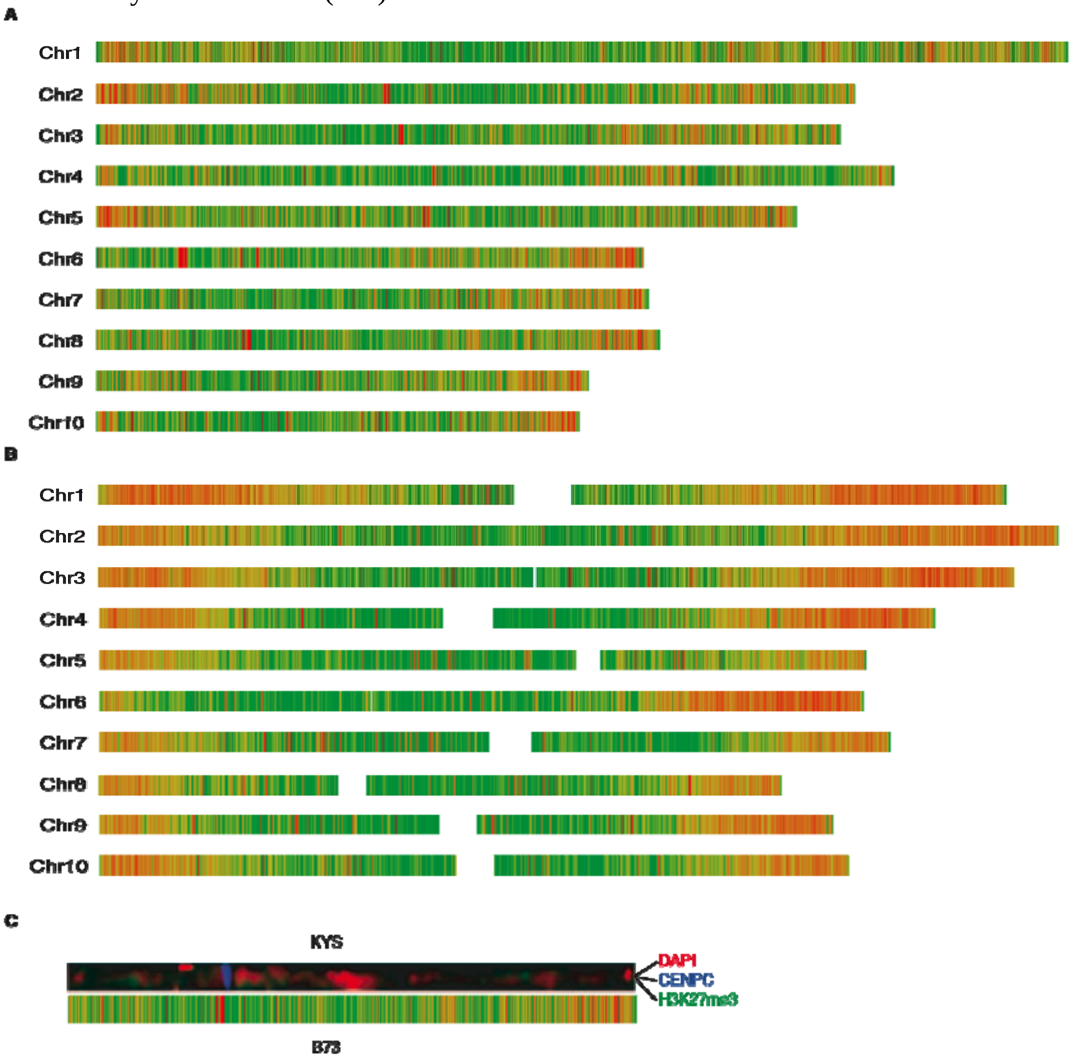


Figure S14. Dot-plots showing maize-rice and maize-sorghum orthologs, classified as not-syntenic (black), syntenic:colinear (blue), and syntenic:in-range (red). See supporting online text for explanation of these terms.

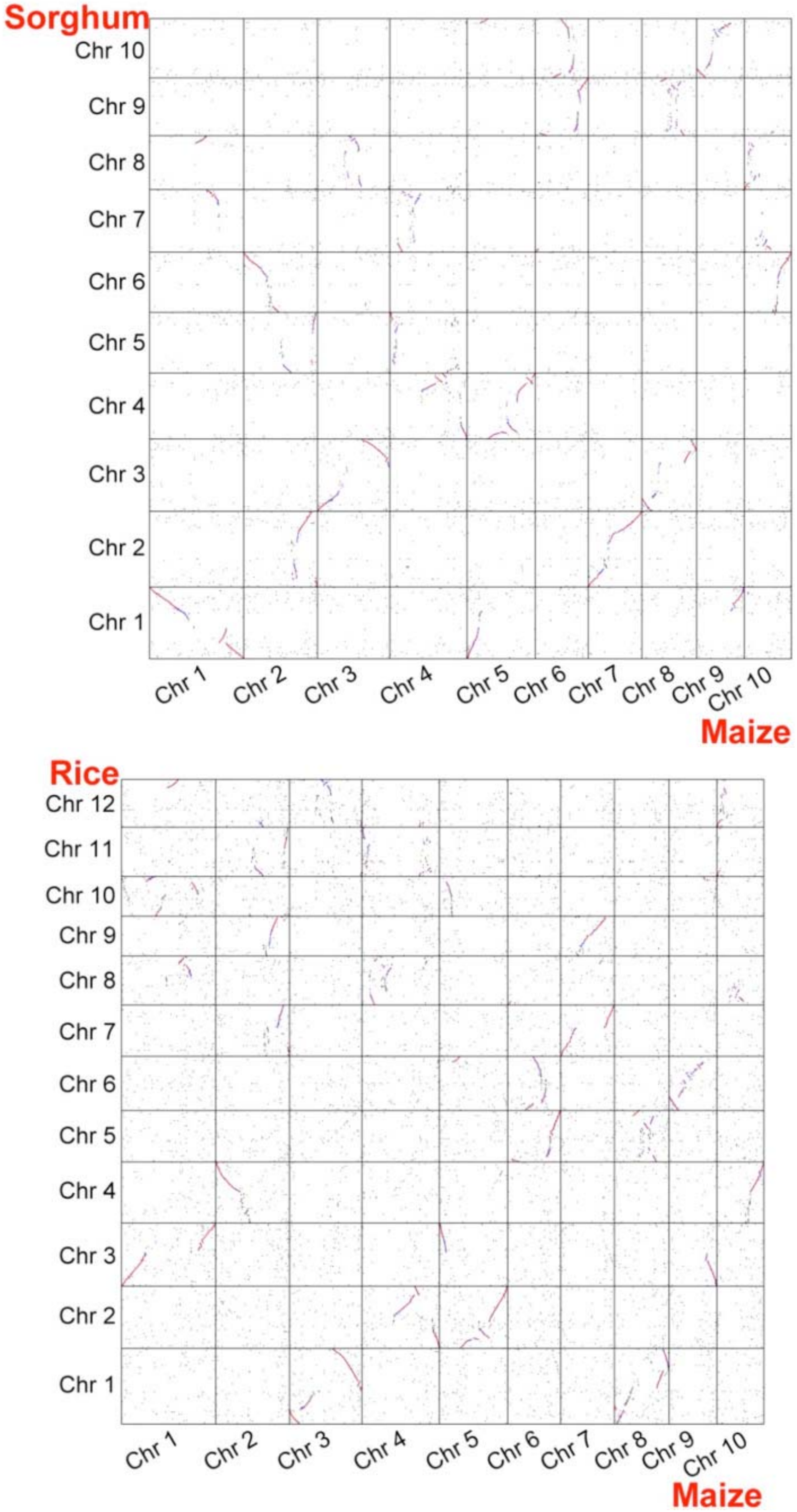


Figure S15. Maize versus maize dot-plot. Black dots show “co-orthologs”, maize genes that have orthology to the same rice or sorghum gene. Red dots show “co-syntenic” maize genes, those that are syntenic to the same rice or sorghum gene.

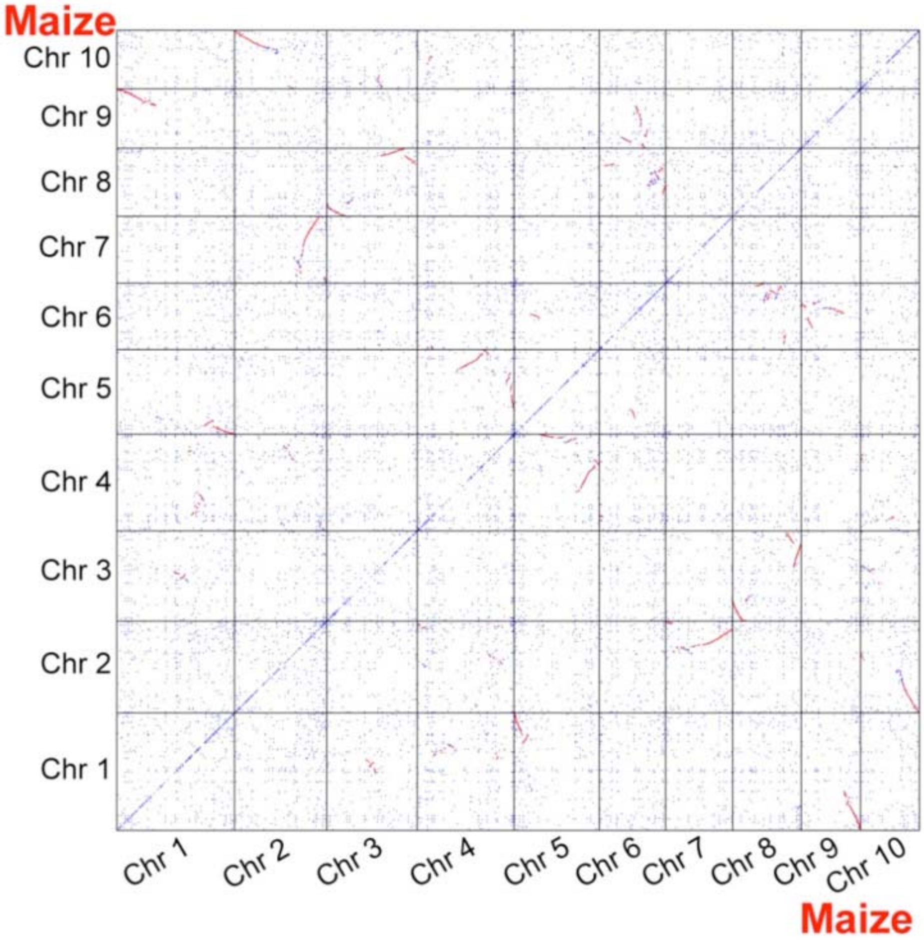


Figure S16. Genes of the CesA/Csl superfamily. At least three distinct cellulose synthase (CesA) genes are co-expressed during primary wall formation and secondary wall formation; mutants in each of them result in cellulose deficiencies, indicating that all three are essential for cellulose synthesis (*S55*). Rice, maize, and sorghum have genes in apparent paralogous clusters with the Arabidopsis CesA genes. Whereas Arabidopsis and rice have ten CesA genes, four additional duplications have occurred in sorghum, and ten in maize. The functions of the cellulose synthase-like (Csl) genes are beginning to be established. The CslA genes are associated with b-mannan synthesis (*S56*, *S57*), CslC genes are involved in xyloglucan formation (*S58*), and some of the CslD genes may be involved in cellulose synthesis in tip-growing cells (*S59*, *S60*). The CslF subfamily of genes is found only in grass species, and heterologous expression indicates its involvement in the synthesis of the grass-specific mixed-linkage (1→3),(1→4)-b-D-glucan (*S61*). In contrast to the CesA family, maize has retained fewer recent duplications in the CslFs than has sorghum. The CslD family comprises five paralogous clusters of each of the grass species, with no recent duplications. The CslA family is greatly expanded in the grasses, with evidence of nine paralogous duplications in grass specific subclades. Grass-specific duplications sometimes resulting in expansion to create new subfamilies is common among the grasses (*S62*). For accession numbers of all genes in this superfamily and other cell wall-related gene families, see http://cellwall.genomics.purdue.edu/families/.

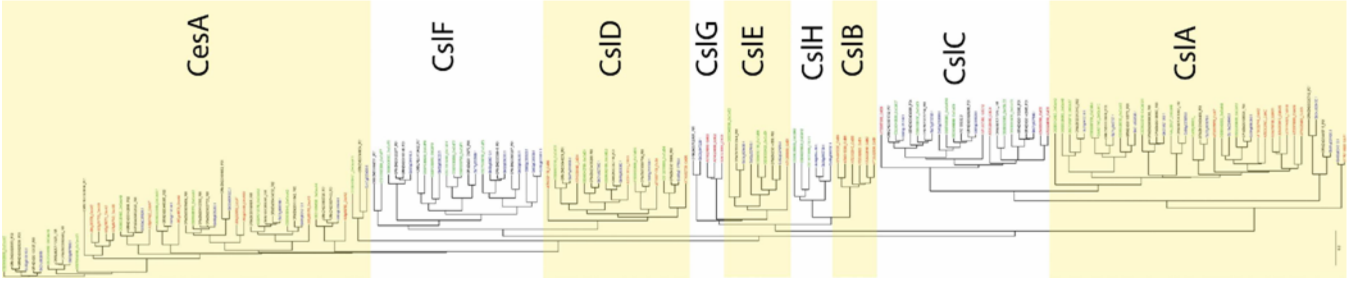


Figure S17. Preferential gene loss between homoeologous maize chromosomes. Percentages of syntenic sorghum genes whose maize syntenic ortholog is present as singletons or retained at both homoeologous sites are shown. Scale is gene index.

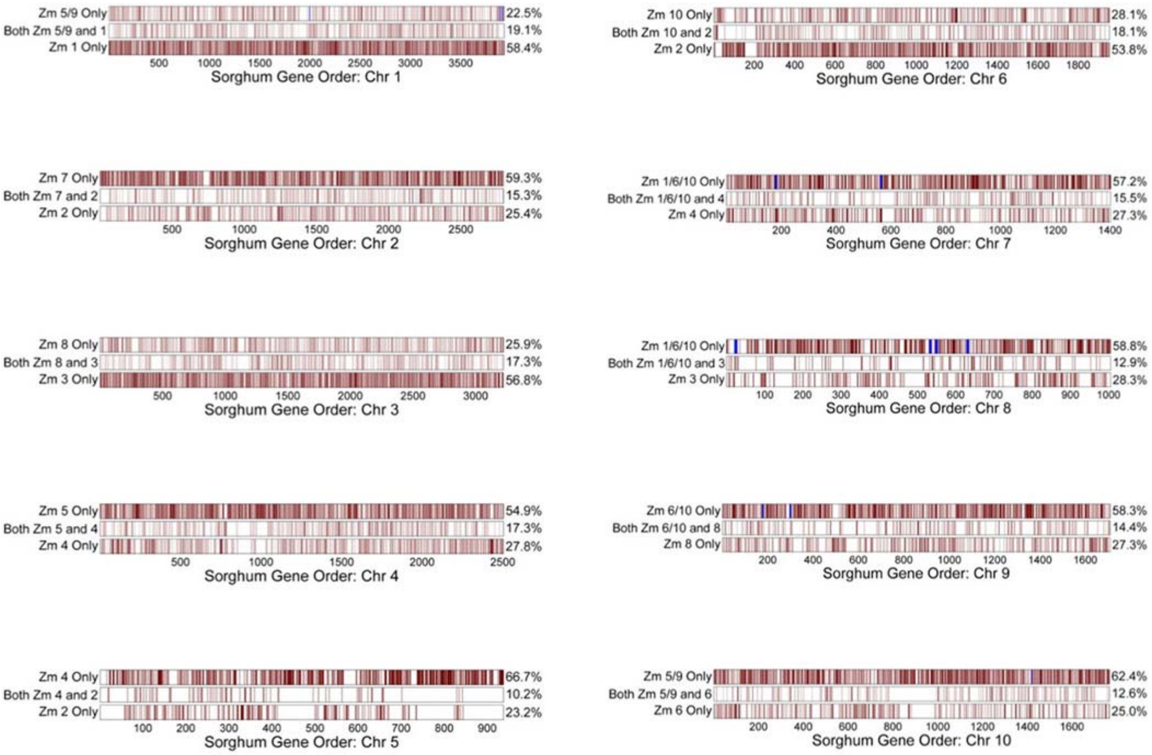
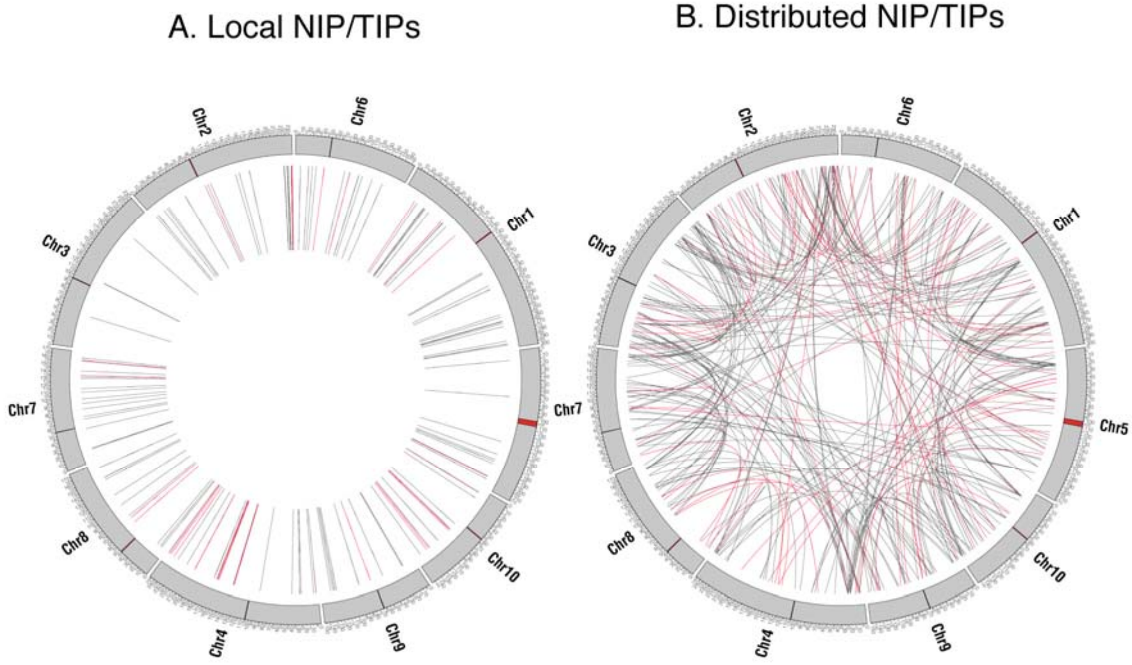


Figure S18. Distribution of NIP/TIP pairs. 222 NIPs exhibit 100% identity; these TIPs (Totally Identical Paralogs) are highlighted in red. (A) Local NIP/ TIP pairs are located within 200 kb of each other; (B) Distributed NIP pairs are >200 kb apart or on different chromosomes. It has been hypothesized that different mechanisms are responsible for the origins of these two classes of NIPs (*S63*). Although NIPs are distributed across the genome, some regions (e.g., 2L and 4S) have elevated rates of inter-chromosomal NIPs. These do not, however, reflect known segment duplication events, arguing against paralog homogenization as a mechanism for the origin of NIPs. On the basis of aCGH experiments ~5% of NIPs have stronger signals in B73 than Mo17 genomic DNA (*S64*), suggesting that Mo17 may have only a single copy of what in the B73 genome are NIP pairs. Centromere positions are from (*S44*).


A. Local NIP/TIPs

B. Distributed NIP/TIPs