The *Sorghum bicolor* genome and the diversification of grasses

Andrew H. Paterson[1]*, John E. Bowers[1], Rémy Bruggmann[2], Inna Dubchak[3], Jane Grimwood[4], Heidrun Gundlach[5], Georg Haberer[5], Uffe Hellsten[3], Therese Mitros[6], Alexander Poliakov[3], Jeremy Schmutz[4], Manuel Spannagl[5], Haibao Tang[1], Xiyin Wang[1,7], Thomas Wicker[8], Arvind K. Bharti[2], Jarrod Chapman[3], F. Alex Feltus[1,9], Udo Gowik[10], Igor V. Grigoriev[3], Eric Lyons[11], Christopher A. Maher[12], Mihaela Martis[5], Apurva Narechania[12], Robert P. Otillar[3], Bryan W. Penning[13], Asaf A. Salamov[3], Yu Wang[5], Lifang Zhang[12], Nicholas C. Carpita[14], Michael Freeling[11], Alan R. Gingle[1], C. Thomas Hash[15], Beat Keller[8], Patricia Klein[16], Stephen Kresovich[17], Maureen C. McCann[13], Ray Ming[18], Daniel G. Peterson[1,19], Mehboob-ur-Rahman[1,20], Doreen Ware[12,21], Peter Westhoff[10], Klaus F.X. Mayer[5], Joachim Messing[2], Daniel S. Rokhsar[3,4]*

[1]Plant Genome Mapping Laboratory, University of Georgia, Athens, GA 30602
[2]Waksman Institute for Microbiology, Rutgers University, Piscataway, NJ 08854
[3]DOE Joint Genome Institute, Walnut Creek, CA 94598
[4]Stanford Human Genome Center, Stanford University Palo Alto, CA 94304
[5]MIPS/IBIS, Helmholtz Zentrum München, Inglostaedter Landstrasse 1, 85764 Neuherberg Germany
[6]Center for Integrative Genomics, University of California, Berkeley, CA 94720
[7]College of Sciences, Hebei Polytechnic University, Tangshan, Hebei 063000, China
[8]Institute of Plant Biology, University of Zurich, Zollikerstrasse 107, 8008 Zurich, Switzerland
[9]Department of Genetics and Biochemistry, Clemson University, Clemson SC 29631
[10]Institut fur Entwicklungs- und Molekularbiologie der Pflanzen, Heinrich-Heine-Universitat, Universitatsstrasse 1, D-40225 Dusseldorf, Germany
[11]Department of Plant and Microbial Biology, University of California, Berkeley 94720
[12]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724
[13]Department of Biological Sciences, Purdue University, West Lafayette, IN 47907
[14]Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN 47907
[15]International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, 502 324, India
[16]Department of Horticulture and Institute for Plant Genomics and Biotechnology, Texas A&M University, College Station, TX 77843
[17]Institute for Genomic Diversity, Cornell University, Ithaca NY 14853
[18]Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801
[19]Mississippi Genome Exploration Laboratory, Mississippi State University, Starkville MS 39762
[20]National Institute for Biotechnology & Genetic Engineering (NIBGE), Faisalabad, Pakistan
[21]USDA NAA Robert Holley Center for Agriculture and Health, Ithaca, New York, 14853

JANUARY 2009

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

The *Sorghum bicolor* genome and the diversification of grasses

Andrew H. Paterson[1]*, John E. Bowers[1], Rémy Bruggmann[2], Inna Dubchak[3], Jane Grimwood[4], Heidrun Gundlach[5], Georg Haberer[5], Uffe Hellsten[3], Therese Mitros[6], Alexander Poliakov[3], Jeremy Schmutz[4], Manuel Spannagl[5], Haibao Tang[1], Xiyin Wang[1,7], Thomas Wicker[8], Arvind K. Bharti[2], Jarrod Chapman[3], F. Alex Feltus[1,9], Udo Gowik[10], Igor V. Grigoriev[3], Eric Lyons[11], Christopher A. Maher[12], Mihaela Martis[5], Apurva Narechania[12], Robert P. Otillar[3], Bryan W. Penning[13], Asaf A. Salamov[3], Yu Wang[5], Lifang Zhang[12], Nicholas C. Carpita[14], Michael Freeling[11], Alan R. Gingle[1], C. Thomas Hash[15], Beat Keller[8], Patricia Klein[16], Stephen Kresovich[17], Maureen C. McCann[13], Ray Ming[18], Daniel G. Peterson[1,19], Mehboob-ur-Rahman[1,20], Doreen Ware[12,21], Peter Westhoff[10], Klaus F.X. Mayer[5], Joachim Messing[2], Daniel S. Rokhsar[3,4]*

* co-corresponding authors

[1]Plant Genome Mapping Laboratory, University of Georgia, Athens, GA 30602
[2]Waksman Institute for Microbiology, Rutgers University, Piscataway, NJ 08854
[3]DOE Joint Genome Institute, Walnut Creek, CA 94598
[4]Stanford Human Genome Center, Stanford University Palo Alto, CA 94304
[5]MIPS/IBIS, Helmholtz Zentrum München, Inglostaedter Landstrasse 1, 85764 Neuherberg Germany
[6]Center for Integrative Genomics, University of California, Berkeley, CA 94720
[7]College of Sciences, Hebei Polytechnic University, Tangshan, Hebei 063000, China
[8]Institute of Plant Biology, University of Zurich, Zollikerstrasse 107, 8008 Zurich, Switzerland
[9]Department of Genetics and Biochemistry, Clemson University, Clemson SC 29631
[10]Institut fur Entwicklungs- und Molekularbiologie der Pflanzen, Heinrich-Heine-Universitat, Universitatsstrasse 1, D-40225 Dusseldorf, Germany
[11]Department of Plant and Microbial Biology, University of California, Berkeley 94720
[12]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724
[13]Department of Biological Sciences, Purdue University, West Lafayette, IN 47907
[14]Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN 47907
[15]International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, 502 324, India
[16]Department of Horticulture and Institute for Plant Genomics and Biotechnology, Texas A&M University, College Station, TX 77843
[17]Institute for Genomic Diversity, Cornell University, Ithaca NY 14853
[18]Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801
[19]Mississippi Genome Exploration Laboratory, Mississippi State University, Starkville MS 39762
[20]National Institute for Biotechnology & Genetic Engineering (NIBGE), Faisalabad, Pakistan
[21]USDA NAA Robert Holley Center for Agriculture and Health, Ithaca, New York, 14853

Summary

Sorghum, an African grass related to sugarcane and maize, is grown for food, feed, fiber, and fuel. We present an initial analysis of the ~730 mbp *S. bicolor* (L.) Moench genome, placing ~98% of genes in their chromosomal context using whole genome shotgun sequence validated by genetic, physical, and synteny information. Genetic recombination is largely confined to about one-third of the sorghum genome with gene order and density similar to those of rice. Retrotransposon accumulation in recombinationally-recalcitrant heterochromatin explains the ~75% larger genome size of sorghum than rice. While gene and repetitive DNA distributions have been preserved since paleopolyploidization ~70 million years ago, most duplicated gene sets lost one member before sorghum/rice divergence. Possible concerted evolution makes one duplicated chromosomal segment appear only a few million years old. About 24% of genes are grass-specific and 7% are sorghum-specific. Recent gene and miRNA duplications may contribute to sorghum's drought tolerance.

The Saccharinae plants (Figure 1) include some of the most efficient biomass accumulators known, providing food and fuel from starch (sorghum) and sugar (sorghum and *Saccharum*, sugarcane), and with promise as cellulosic biofuel crops (sorghum, sugarcane, *Miscanthus*). Of singular importance to the productivity of Saccharinae grasses is 'C4' photosynthesis, comprising biochemical and morphological specializations that increase net carbon assimilation at high temperatures[1]. The Saccharinae exhibit much morphological, physiological, and genome size variation, both polyploidization and chromosome number reduction, and introgression across several species boundaries (Supplementary Figure 1).

Its small genome (~730 Mb) makes sorghum an attractive model for functional genomics of Saccharinae and other grasses using C4 photosynthesis. Rice, with the first fully sequenced cereal genome, is more representative of C3 photosynthetic grasses. Drought tolerance makes sorghum especially important in dry areas such as Northeast Africa (its center of diversity) and the US Southern Plains. Genetic variation in perenniality, as well as in partitioning of carbon into sugar stores versus cell wall mass and associated physiological and architectural features such as tillering and stalk reserve retention[2] make sorghum an attractive system for study of many traits important in perennial cellulosic biomass crops.

Assembling a retrotransposon-rich plant genome

Preferred approaches to sequence entire genomes are currently to apply shotgun sequencing[3] either to a minimum 'tiling path' of genomic clones, or to genomic DNA directly. The latter approach, whole genome shotgun sequencing (WGS), is widely used for mammalian genomes, being fast, relatively economical, and reducing cloning bias. However, its applicability has been questioned for repetitive DNA-rich plant genomes[4].

Despite ~61% repeat content, a high quality sorghum genome sequence was assembled from homozygous genotype BTx623 by using WGS and incorporating two cardinal features: (1) ~8.5 genome-equivalents of paired-end reads[5] from genomic libraries spanning a ~100-fold range of insert sizes (Table S1) resolved many repetitive regions; and (2) average high-quality read length of 723 bp facilitated assembly. Divergence among many members of repetitive element 'families' was sufficient to allow their disambiguation, accurately reconstructing large genomic regions. Comparison with 27 finished BACs that sample diverse genomic regions showed the WGS assembly to be both complete (>98.46%) and accurate (<1 error/10 kb: Supplementary Note 2.5).

Comparison of the WGS assembly with a high-density genetic map[6], an FPC-based physical map richly populated with sequence-tagged probes[7], and the rice sequence[4] helped to reconstruct the sorghum genome (Supplementary Notes 1-2). The 201 largest WGS scaffolds span 678.9 mbp and represent 97.3% of the assembly. A total of 28 assembly errors in these scaffolds were identified based on discrepancies with the genetic and/or physical maps, each supported by multiple lines of evidence (Supplementary Note 2.6) and often involving repetitive elements. A total of 38 (2%) of 1869 FPC contigs[7] were deemed erroneous, containing >5 BAC-ends that fell into different sequence scaffolds. After breaking the WGS assembly at the 28 points of discrepancy, the resulting 229 scaffolds have N50 of 35 and L50 of 7.0 Mb.

A total of 127 scaffolds containing 625.7 mbp (89.7%) of DNA and 1,476 FPC contigs could be assigned to chromosomal locations and oriented based on physical map, genetic map, rice synteny, genome structure (gene and repeat distributions), and cytological information[8]. The other 102 scaffolds were generally smaller (53.2 mbp, 7.6% of nucleotides) and heterochromatic, with only 374 predicted genes and 85 (83%) scaffolds containing large stretches comprised predominantly of the CEN38[9] centromeric repeat. These 102 scaffolds merged only 193 FPC contigs, presumably due to the greater abundance of repeats that are recalcitrant to clone-based physical mapping [7] and may be omitted in BAC-by-BAC approaches [10]. Most chromosomal models appeared largely complete – 15 of 20 terminated in telomeric repeats (Supplementary Note 2.3).

Genome size evolution and its causes

The ~75% larger quantity of DNA in the genome of sorghum than rice is mostly heterochromatin. Alignment to genetic[6] and cytological maps[8] suggests that sorghum and rice have similar quantities of euchromatin (252 and 309 mbp: Supplementary Table 7). Euchromatin accounts for 97-98% of recombination (1025.2 cM and 1496.5 cM) and 75.4-94.2% of genes in the respective cereals, with largely collinear gene order[7]. In contrast, pericentromeric heterochromatin occupies at least 460 mbp (62%) in sorghum versus 63 mbp (15%) in rice, and may be underestimated because of its recalcitrance to clone-based physical mapping[7] in the rice BAC-based sequence[4] and to assembly in the sorghum WGS sequence. The ~3x genome expansion in maize since its divergence from sorghum[11] has been more dispersed –highly recombinogenic DNA has grown to ~1382 mbp, a much greater increase (4.5x) than can be explained by its genome duplication[12].

The net size expansion of the sorghum genome relative to rice largely involved LTR-retrotransposons. The sorghum genome contains 55% retrotransposons, intermediate between the ~3x larger maize genome (79%) and the rice genome (26%). However, sorghum more closely resembles rice in having a higher ratio of *gypsy*- to *copia*-like elements (3.7 to 1 and 4.9 to 1) than maize (1.6 to 1: Supplementary Table 10).

While recent retroelement activity is widely distributed across the sorghum genome, turnover is rapid (as in other cereals[13]) with pericentromeric elements persisting longer. Very recent insertions of LTR retrotransposons (<0.01 mya) appear randomly distributed across the chromosomes, suggesting that they are preferentially eliminated from gene-rich regions [7] but more free to accumulate in gene-poor regions (Figure 2; Supplementary Note 3.1). LTR-retrotransposon insertion times for one representative sorghum chromosome, 8, suggest a major wave of retrotransposition less than 1 mya, following a smaller wave 1-2 mya (Figure S2).

CACTA-like elements, the predominant class of sorghum DNA transposon (4.7% of the genome), appear to relocate genes and gene fragments. Mutator-like 'Pack-MULE' elements are important gene-transducing elements[14] in rice, and intact helitrons are implicated in maize gene movement[15]. Among 95 novel CACTA families discovered in sorghum, most individual elements are non-autonomous deletion derivatives in which the typical transposon genes have been replaced with non-transposon DNA including exons from one or more genes. For example, CACTA family *G118* (Figure 3) has only one complete and presumably autonomous "mother" element. Among 18 deletion derivatives, only the terminal 500-2500 bp are conserved, with 8 carrying gene fragments internally. One relatively homogeneous subgroup (G118_106, 111 and 112) presumably arose recently, while all other derivatives are unique. Among the 13,775 CACTA elements identified (Supplementary Note 3.4), 200 encode no transposon proteins but contain at least one fragment of a cellular gene. The actual number of CACTA-vectored gene fragments might be significantly higher because many CACTA elements are truncated, making it difficult to determine whether nearby genes were vectored or native.

In total, DNA transposons constitute 7.5% of the sorghum genome, intermediate between maize (2.7%) and rice (13.7%: Supplementary Table 10). Miniature inverted-repeat transposable elements (MITEs) are 1.7% of this, and are closely associated with genes (Fig. 2; Supplementary Note 3) as in other cereals[16]. Helitrons comprise ~0.8% of the sorghum genome, nearly all lacking helicase as is true of most maize helitrons[15], but with possible gene fragments inferred (Supplementary Note 3.5). Helitrons carrying genes or gene fragments appear more abundant in maize than sorghum with 1.3% detected in 100 randomly selected BACs and 1.8% (Supplementary Table 1$) in two large contiguous genomic sequences[17,18]. The latter regions are gene-rich indicating that helitrons are more abundant in such areas.

Organellar DNA insertion has contributed only about 0.085% to the sorghum nuclear genome, far less than the 0.53% of rice. Organellar DNA shows more sequence conservation with longer nuclear insertions, suggesting that they are more prone to removal than short insertions (Supplementary Note 2.7).

The gene complement of sorghum

Among 34,496 sorghum gene models, we found ~27,640 *bona fide* protein-coding genes by combining homology-based and *ab initio* gene prediction methods with expressed sequences from sorghum, maize, and sugarcane (Supplementary Note 4). Evidence for alternate splicing is found in 1,491 loci.

Another 5,197 predicted gene models are typically shorter than the *bona fide* genes (often <150 amino acids); have few exons (often one) and no EST support (vs. 85% for *bona fide* genes); are more diverged from related rice genes; and are often found in large families enriched for "hypothetical," "uncharacterized," and/or retroelement-associated domains and annotations, despite repeat masking of the genome (Supplementary Note 4). Relatively high concentration in the pericentromeric regions where *bona fide* genes are scarce (Fig. 2) suggests that many of these low confidence gene models are retroelement-derived. We also identified 727 processed pseudogenes and 932 predictions containing domains known only from transposons.

The exon size distribution of orthologous sorghum and rice genes shows nearly perfect agreement, and intron position and phase show >98% concordance (Supplementary Note 5). Conserved intron position and phase between *Arabidopsis* and rice[19] extend the conservation of gene structure back to the last common eudicot-monocot ancestor. Even intron size has been highly conserved between sorghum and rice, although it has increased in maize due to transpositions[17].

Most paralogs in sorghum are proximally duplicated, including 5,303 genes in 1,947 families of two or more genes. (Supplementary Note 4.3). The longest tandem gene array is 15 cytochrome P450 genes. Other sorghum-specific tandem gene expansions (3 or more) include haloacid dehalogenase-like hydrolases (PF00702); FNIP repeats (PF05725), and male sterility proteins (PF03015).

We confirmed the genomic locations of 67 known sorghum miRNAs and identified 82 additional miRNAs (Supplementary Note 4.4). Five clusters located within 500bp of each other represent putative polycistronic miRNAs, similar to those in *Arabidopsis* and *Oryza*. Natural antisense miRNA precursors (nat-miRNAs) of families miR444 [20] have been identified in three copies. One sbi-miR444 locus produces two precursors, due to exon skipping.

Comparative gene inventories of angiosperms

The number and sizes of sorghum gene families are similar to those of *Arabidopsis*, rice and poplar (Figure 4: Supplementary Note 4.6). A total of 9,503 (58%) sorghum gene families were shared among all four species and 15,225 (93%) overlapped with at least one other species. Nearly 94% of high confidence sorghum genes (25,875/27,640) have orthologs in rice, *Arabidopsis*, and/or poplar, and together these gene complements

define 11,502 ancestral angiosperm gene families represented in at least one contemporary grass and rosid genome. However, 3,983 (24%) gene families have members only in the grasses sorghum and rice; and 1,153 (7%) appear unique to sorghum. A similar percentage of unique gene families is observed for *Arabidopsis* (6.7%), with fewer in rice (3.6%) and more in poplar (15.7%).

PFAM domains that are over-represented, under-represented or even absent in sorghum relative to rice, poplar and *Arabidopsis*, may reflect biological peculiarities specific to the *Sorghum* lineage. Domains over-represented in sorghum are usually present in the other organisms, a notable exception being the alpha kafirin domain that accounts for most sorghum seed storage protein (Supplementary Table 20). The kafirin genes are absent from rice, but correspond to maize zeins[21]. The kafirins have propagated proximally, with at least 14 copies within a megabase-sized segment of sorghum chromosome 5.

NBS-LRR containing proteins associated with the plant immune system are only about half as frequent in sorghum as in rice. A search of with 12 NBS domains from published rice, maize, wheat and *Arabidopsis* NBS-LRR gene sequences revealed 211 NBS-LRR coding genes in sorghum, versus 410 in rice, and 149 in *Arabidopsis*[22]. Sorghum NBS-LRR genes mostly encode the CC type of N-terminal domains. Only two sorghum genes (Sb02g005860, Sb02g036630), annotated as TIR-P-loop LRR genes, contain the TIR domain, and neither contains an NBS domain. NBS-LRR genes are most abundant on sorghum chromosome 5 (62), and its rice homolog (chromosome 11, 106 NBS-LRR genes). Enrichment of NBS-LRR genes in particular genomic regions may suggest evolution of R gene location, in contrast to a proposal that gene movement would be specifically advantageous for R genes[23].

Evolution of distinctive pathways and processes

The evolution of C_4 photosynthesis in the sorghum lineage involved redirection of C_3 progenitor genes as well as recruitment and functional divergence of both ancient and recent gene duplicates. The sole sorghum C_4 pyruvate orthophosphate dikinase (*ppdk*) and the phosphoenolpyruvate carboxylase kinase (*ppck*) gene and its two isoforms (produced by the whole genome duplication) have only single orthologs in rice. Additional duplicates formed in maize after the sorghum-maize split (*Zm-ppck*2 and *Zm-ppck*3). The C4 NADP dependent malic enzyme (*me*) gene has an adjacent isoform but each corresponds to a different maize homolog, suggesting tandem duplication before the sorghum-maize split. The C_4 malate dehydrogenase (*mdh*) gene and its isoform are also adjacent, but share 97% amino acid similarity and correspond to the single known maize *mdh* gene, suggesting tandem duplication in sorghum after its split with maize. The rice *me* and *mdh* genes are single copy, suggesting duplication and recruitment to the C_4 pathway after the Panicoideae-Oryzoideae divergence. See Supplementary Note 9 for further details.

The sorghum sequence reinforces inferences previously based only on rice, about how different grass and dicot gene inventories may relate to their two distinct types of cell

walls[24]. About 2500 genes in 80 families function in cell wall biogenesis. In grasses, cellulose microfibrils coated with mixed-linkage (1→3),(1→4)-β-D-glucans are interlaced with glucuronoarabinoxylans and an extensive complex of phenylpropanoids[25]. The sorghum sequence largely corroborates differences between dicots and rice in the distribution of genes within some of the gene families (Supplementary Note 10). For example, the CesA/Csl superfamily and callose synthases have either diverged so significantly as to form new sub-groups or functionally non-essential sub-groups were selectively lost, such as *CslB* and *CslG* lost from the grass species, and *CslF* and *CslH* lost from species with dicot-like cell walls[26]. The previously rice-unique *CslF* and *CslH* genes are present in sorghum. *Arabidopsis* contains a single Group F GT31 gene, whereas sorghum and rice contain six and ten members, respectively. The protein sequence relatedness and clustering of genes along three chromosomal regions in rice and two in sorghum suggests that they have arisen from recent duplication events after the grass/dicot split.

The characteristic adaptation of sorghum to drought may be partly related to expansion of one miRNA and several gene families. Rice miRNA 169g, up-regulated during drought stress [27], has five sorghum homologs (sbi-MIR169c&d, sbi-MIR169.p2, sbi-MIR169.p6 and sbi-MIR169.p7). The computationally predicted target of the sbi-MIR169 subfamily comprises members of the plant nuclear factor Y (NF-Y) B transcription factor family, linked to improved performance under drought for both *Arabidopsis* and maize[28]. Cytochrome P450 domain-containing genes, often involved in scavenging toxins such as those accumulated in response to stress, are also unusually abundant in sorghum with 326 family members versus only 228 in rice. With 82 copies in sorghum versus 58 in rice and 40 each in *Arabidopsis* and poplar. another large gene family that could be linked to the durability of sorghum is the expansins, enzymes that break hydrogen bonds and are responsible for a variety of plant growth responses.

Duplication and diversification of cereal genomes

Whole-genome duplication in a common ancestor of cereals is reflected in 'quartet' alignments (Figure 5) of sorghum and rice genes. Among 34,496 non-transposon sorghum gene models, 19,929 (57.8%) were in blocks collinear with rice (Supplementary Note 6). A total of 13,667 (68.6%) of the colinear genes retained only one copy following the whole-genome duplication, with 13,526 (99%) being orthologous in rice-sorghum, suggesting that most gene losses predate their divergence. Both sorghum and rice retained both copies of 4912 (14.2%) genes, while sorghum lost one copy of 1070 (3.1%) and rice lost one copy of 634 (1.8%). These patterns are likely to be predictive of other cereal genomes, since the major cereal lineages are thought to have diverged from a common ancestor about the same time[29] (see also Supplementary Note 7).

While most post-duplication gene loss happened in a common cereal ancestor, some lineage-specific patterns occur. A total of 2 and 10 protein functional (Pfam) domains showed enrichment for duplicates and singletons (respectively) in sorghum but not rice (Supplementary Note 6.1). Since sorghum-rice divergence is thought to have been 20 my or more after the genome duplication[29], this suggests that even long-term gene loss

is not random but differentially affects gene functional groups. Future revision of inferred gene retention/loss patterns[30] to consider sorghum-rice synteny will reduce artifacts, for example distinguishing cases in which a gene recently migrated to a locus from those in which an ancestral duplicate was lost.

One genomic region has been subject to a high level of concerted evolution. It was previously suggested that rice chromosomes 11 and 12 share a segmental duplication near the termini of the short arms, dated to ~5-7 mya[31]. We found a duplicated segment in the corresponding regions on the orthologous sorghum chromosomes, 5 and 8. Sorghum-sorghum and rice-rice paralogs from this region show Ks values of 0.44 and 0.22 respectively, consistent with only 34 and 17 my of divergence. However, sorghum-rice orthologs show a Ks of 0.63, similar to the genome wide averages for sorghum (0.81) and rice (0.87). We suggest that the sorghum 5-8 (= rice 11-12) duplication resulted from the pan-cereal whole-genome duplication and became differentiated from the remainder of the chromosome(s) due to concerted evolution acting independently in sorghum, rice, and perhaps other cereals. Gene conversion and illegitimate recombination are more frequent in the rice 11-12 region than anywhere else in the genome[32]. Physical and genetic maps suggest shared terminal segments of the corresponding chromosomes in wheat (4, 5), foxtail millet (VII, VIII), and pearl millet (linkage groups 1, 4)[33].

Synthesis and implications

Comparison of the sorghum and rice genomes with one another and other genomes clarifies the cereal gene set. Pairs of orthologous sorghum and rice genes, combined with recent paralogous duplications in each genome, define 19,542 conserved grass gene families, each representing a single gene in the sorghum-rice common ancestor. While our sorghum gene count is similar to the number in a manually curated rice annotation (RAP2)[34], this similarity masks some differences among these annotations and the automated TIGR5 annotation[35]. About 2054 syntenic orthologs shared by our sorghum annotation and TIGR5 are absent from RAP2. Conversely, ~12,000 TIGR5 annotations may be transposable elements or pseudogenes, based on their presence in large families of hypothetical genes in both sorghum and rice, and/or short coding length, small intron number, and limited EST support. Phylogenetically-incongruent patterns of apparent gene retention/loss in these and other taxa (for example, genes shared by *Arabidopsis* and sorghum but not rice: Figure 4) may also suggest misannotations.

Comparison of sorghum and rice underlines the bipolar nature of angiosperm genomes. Synteny is highest and retroelement abundance lowest in distal portions of the chromosomes. Despite nearly complete turnover of specific elements, patterns of repetitive DNA organization have been substantially preserved since the divergence of chromosomes that duplicated 70 mya, remaining correlated in paleo-duplicated chromosomes (Fig. 2). More rapid removal of retroelements from gene-rich euchromatin (which frequently recombines) than pericentromeric heterochromatin (which rarely recombines), supports the hypothesis that recombination may preserve gene order by exposing new rearrangements to selection[7]. Less polarization in maize,

where retrotransposon persistence in euchromatin appears more frequent, may reflect variation in organization patterns of different cereal genomes or perhaps a lingering consequence of maize genome duplication.

Conserved sequences, both coding and noncoding, among maximally diverged cereal genomes may help us understand the essential genes and binding sites that define grasses. Progress in sequencing of *Brachypodium distachyon*[36] sets the stage for panicoid-oryzoid-pooid phylogenetic triangulation of genomic changes, as well as identification of associations between these changes and phenotypes ranging from molecular (gene expression patterns) to morphological. The divergence between sorghum and either rice or *Brachypodium* is sufficient to randomize nonfunctional sequence and permit conserved noncoding sequence (CNS) discovery by DNA sequence alignment[37] (Figure S9). More distant comparisons such as to the dicot *Arabidopsis* show exon conservation but no CNSs (Figure S10). Chloridoid and arundinoid sequences are needed to sample the remaining cereal lineages, including additional food, turf, forage, and biofuel crops. The sequence of a cereal outgroup such as *Ananas* (pineapple) or *Musa* (banana) would further aid in identifying genes and sequences that define cereals.

The fact that the sorghum genome has not re-duplicated since the ~70 mya cereal duplication[29] makes it a valuable outgroup for deducing the fates of gene pairs and CNS following more recent duplications in related grasses. Individual sorghum regions correspond to two distinct regions resulting from maize-specific genome doubling [38] -- gene fractionation is evident (Figure 5), and subfunctionalization is probable (Figure S10). Sorghum may prove even more valuable for deducing the consequences of additional genome duplications in the more closely-related *Saccharum-Miscanthus* clade; Sugarcane has undergone at least two genome duplications since its divergence from sorghum 8-9 mya[39] and the resulting polyploidy and heterozygosity complicate its genetics[40] yet *Saccharum* BACs show substantially conserved gene order with sorghum (Supplementary Note 11).

Strong conservation of gene structure and colinearity among other cereals facilitates the development of DNA markers to support crop improvement. We identified about 71,000 SSRs in sorghum (Supplementary List 1); among a sampling of 212, only 9 (4.2%) map to a paralog of their source locus. Conserved-intron scanning primers (CISPs: Supplementary List 2) for 6,760 genes provide DNA markers useful across many Poaceae and even non-Poaceae monocots, particularly valuable for 'orphan cereals' that lack maps[41].

As the first plant genome of African origin to be sequenced, sorghum adds new dimensions to ethnobotanical studies. Of particular interest will be the identification of genes (alleles) related to the earliest stages of sorghum cultivation, and a test of the hypothesis that convergent mutations in corresponding genes may have contributed to independent domestications of divergent cereals on different continents[42]. Invigorated sorghum improvement would particularly benefit regions such as the West African

'Sahel' where drought tolerance makes sorghum a staple for human populations that are increasing by 2.8% per year while sorghum yields only gained a total of 6% from 1961-1963 to 1999-2001[43].

Acknowledgements

We thank the US Department of Energy Joint Genome Institute Community Sequencing Program for sequencing sorghum, especially Jim Bristow, Susan Lucas and the JGI production sequencing team; and L. Lin for contributions to Figure 1. We appreciate funding from the US National Science Foundation (NSF DBI-9872649, 0115903; MCB-0450260), International Consortium for Sugarcane Biotechnology, National Sorghum Producers, and a John Simon Guggenheim Foundation fellowship to AHP; US Department of Energy (DE-FG05-95ER20194) to JM; German Federal Ministry of Education in the frame of the GABI initiative to MIPS (0313117 and 0314000C); NSF DBI-0321467 to AN; and US Department of Agriculture-Agricultural Research Service to CA, LZ, and DW.

Figure 1. **Evolutionary context of sorghum.** Branch lengths above the species level were computed by aligning EST assemblies from the TIGR PlantTA collection (plantta.tigr.org), estimating the transversion rate at fourfold synonymous sites using a Jukes-Cantor correction for multiple transversions, and creating a phylogenetic tree with the neighbor-joining method implemented in Phylip (evolution.genetics.washington.edu/phylip.html).

Figure 2: Genomic landscape of sorghum chromosomes 3 and 9. Area charts show the abundance of the four main DNA element types constituting the sorghum genome: retrotransposons (55%), genes (6% exons, 8% introns), DNA transposons (7%) and centromeric repeats (2%). The as-yet unassigned (gray) portion of the genome includes regulatory regions. Alignment of chromosomes 3 and 9 is shown by lines connecting corresponding duplicated genes. Heatmap tracks provide greater detail regarding the distribution of selected elements. Gene densities are highest near chromosome ends and retrotransposon abundance is highest in pericentromeric space, with a gradual and discontinous transition. The LTR-copia retrotransposon superfamily is more widely-distributed than the gypsy superfamily. MITE DNA transposons are gene-associated while CACTA elements are widespread but with hotspots in gene-poor regions. Figures for all 10 sorghum chromosomes are provided (Supplementary Note 3). Abbreviations: Cen38: sorghum specific centromeric repeat[9]; RTs: retrotransposons (class I); LTR-RTs: Long terminal repeat retrotransposons; DNA-TEs: DNA transposons (class II); hc genes: high confidence genes.

Figure 3: CACTA element deletion derivatives that carry gene fragments. The locations of the hits to known rice proteins are indicated as coloured boxes. The descriptions of the foreign gene fragments are indicated underneath the boxes. (HP = Hypothetical protein).

Figure 4: Orthologous gene families between sorghum, *Arabidopsis*, rice and poplar. The numbers of gene families (clusters) and the total numbers of clustered genes are indicated for each species and species intersection.

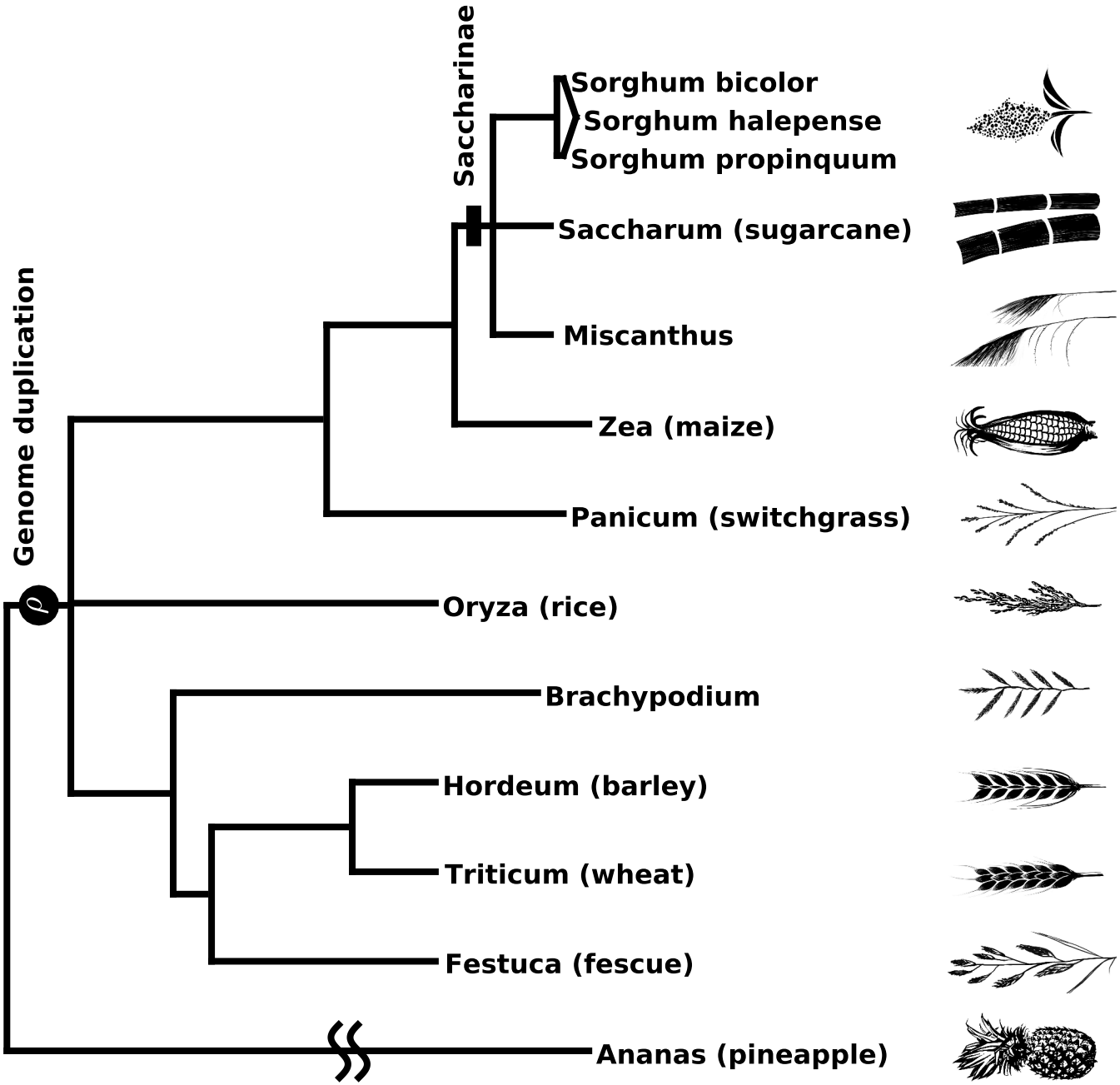
Figure 5: Multi-alignment of corresponding genomic regions of sorghum, rice, and maize. Sorghum and rice form collinear quartets, with two paralogous regions within each genome derived from whole-genome duplication in a common ancestor (see Supplementary Materials; for gene accessions, see quartet ID 03-1322 to 03-1367. Genome-wide dot-plot-based alignments are in Supplementary Note 6). Sorghum-rice orthologs are more similar than rice-rice paralogs, although infrequent gene loss following sorghum/rice divergence causes 'special cases' in which there is a paralog resulting from whole-genome duplication but no ortholog. For illustration, the putative site of the missing gene is interpolated as the middle of flanking collinear gene pairs. Each sorghum region corresponds to two distinct maize regions formed by genome doubling following sorghum-maize divergence[38]. Since most maize BACs are not yet finished we connect syntenic pairs from sorghum loci to the centers of appropriate maize BACs. Note the different scale necessary for maize physical distance.

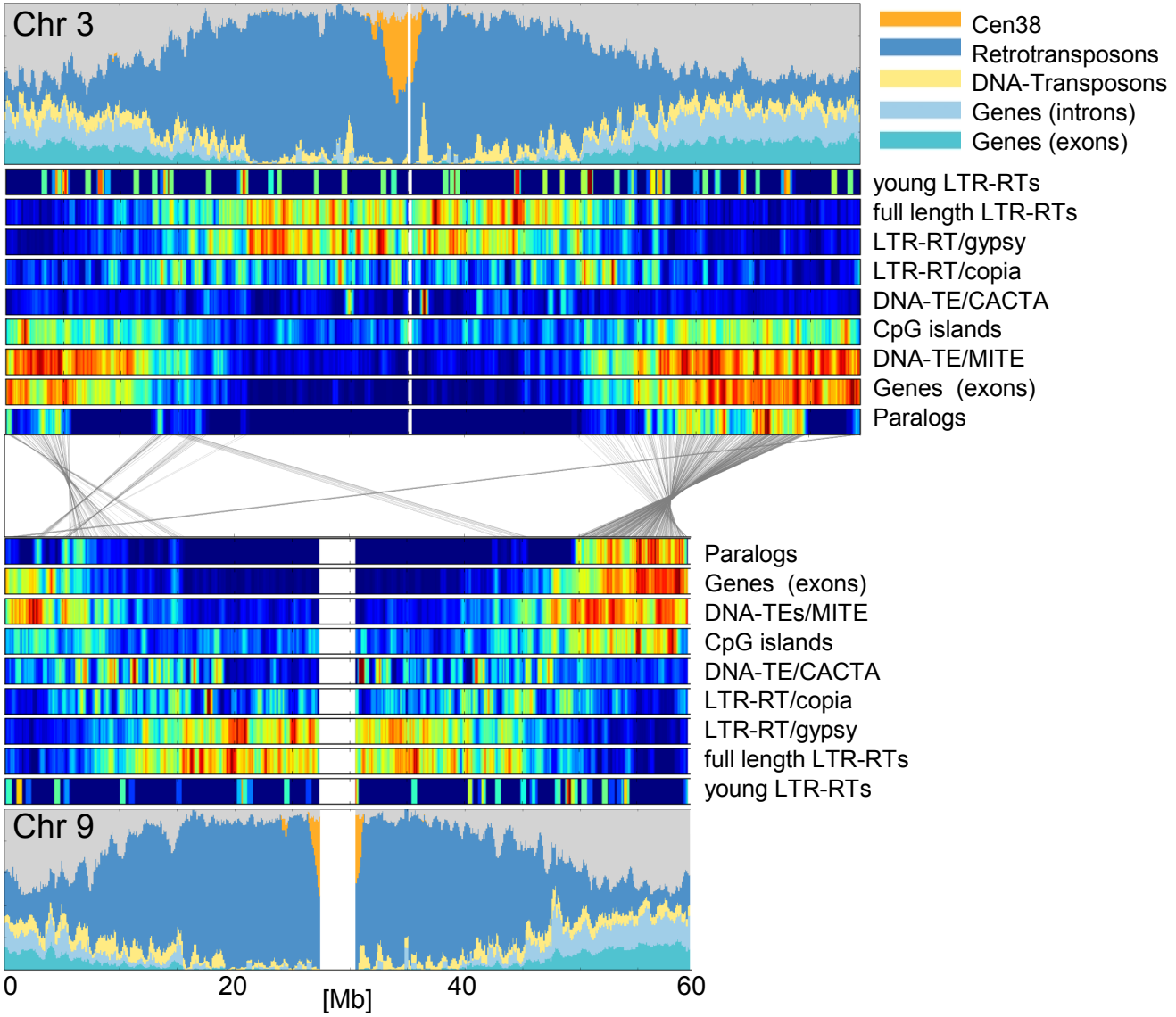
Figure 6: Independent illegitimate recombination in corresponding regions of sorghum and rice. Four homoeologous rice and sorghum chromosomes (R11, R12, S5, S8) are shown, with gene densities plotted. 'L' and 'S' show long and short arms. Lines show Ks between homoeologous gene pairs, and colors are used to show different dates of conversion events.

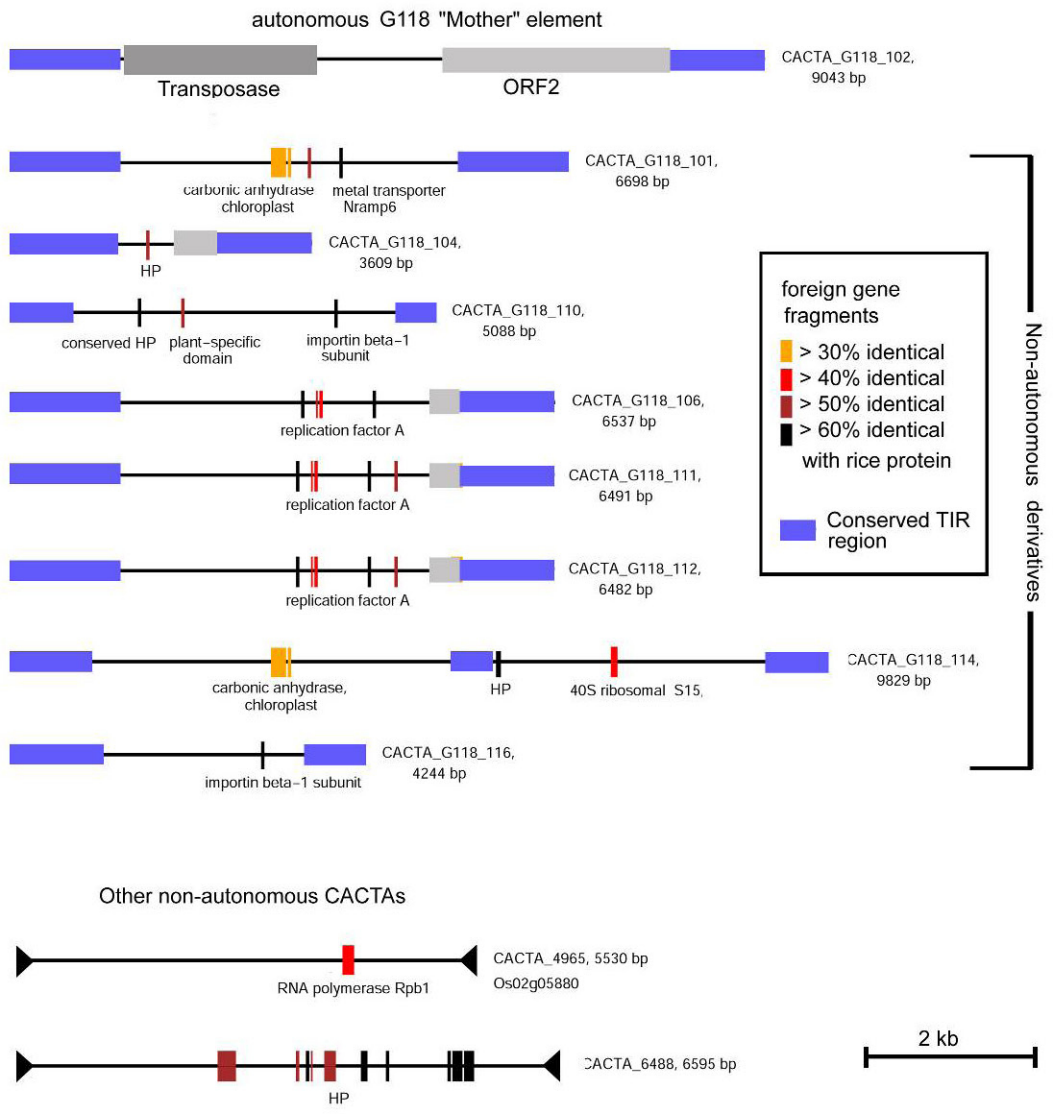
REFERENCES

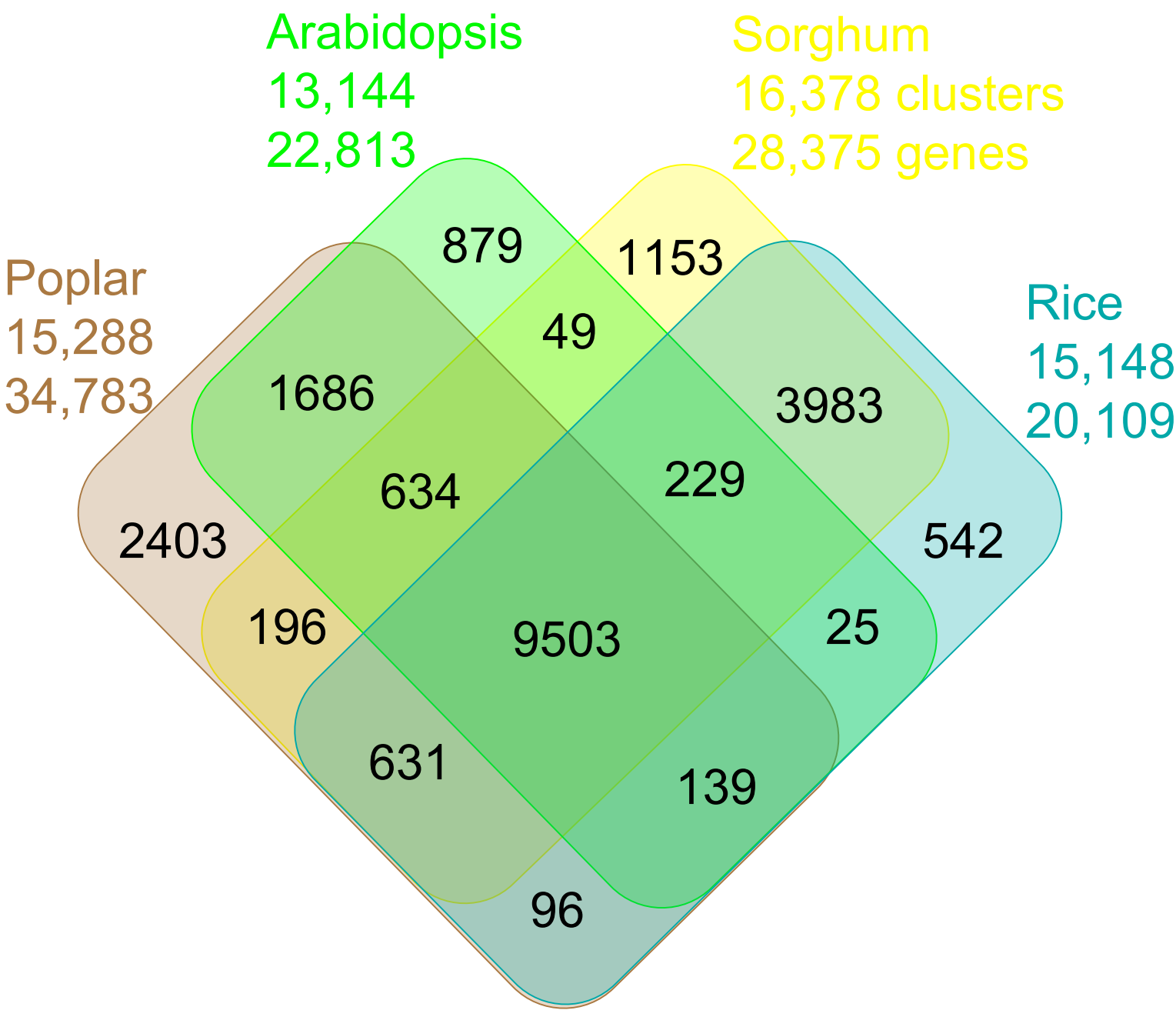
1 M.D. Hatch and C.R. Slack, *Biochemical Journal* **101** (1), 103 (1966).
2 A. H. Paterson, K. F. Schertz, Y. R. Lin et al., *Proceedings of the National Academy of Sciences of the United States of America* **92** (13), 6127 (1995).
3 R. C. Gardner, A. J. Howarth, P. Hahn et al., *Nucleic Acids Research* **9** (12), 2871 (1981).
4 T. Matsumoto, J. Z. Wu, H. Kanamori et al., *Nature* **436** (7052), 793 (2005).
5 J. Vieira and J. Messing, *Gene* **19** (3), 259 (1982).
6 J. E. Bowers, C. Abbey, S. Anderson et al., *Genetics* **165**, 367 (2003).
7 J. E. Bowers, M. A. Arias, R. Asher et al., *Proceedings of the National Academy of Sciences of the United States of America* **102** (37), 13206 (2005).
8 J. S. Kim, P. E. Klein, R. R. Klein et al., *Genetics* **169** (2), 1169 (2005).
9 J. T. Miller, S. A. Jackson, S. Nasuda et al., *Theoretical and Applied Genetics* **96** (6-7), 832 (1998).
10 J. C. Venter, M. D. Adams, G. G. Sutton et al., *Science* **280** (5369), 1540 (1998).
11 Z. Swigonova, J. Lai, J. Ma et al., *Genome Research* **14**, 1916 (2004).
12 Z. Swigonova, J. S. Lai, J. X. Ma et al., *Comparative and Functional Genomics* **5** (3), 281 (2004).
13 Z. Swigonova, J. L. Bennetzen, and J. Messing, *Genetics* **169** (2), 891 (2005).
14 N. Jiang, Z. R. Bao, X. Y. Zhang et al., *Nature* **431** (7008), 569 (2004).
15 S. Brunner, K. Fengler, M. Morgante et al., *Plant Cell* **17** (2), 343 (2005).

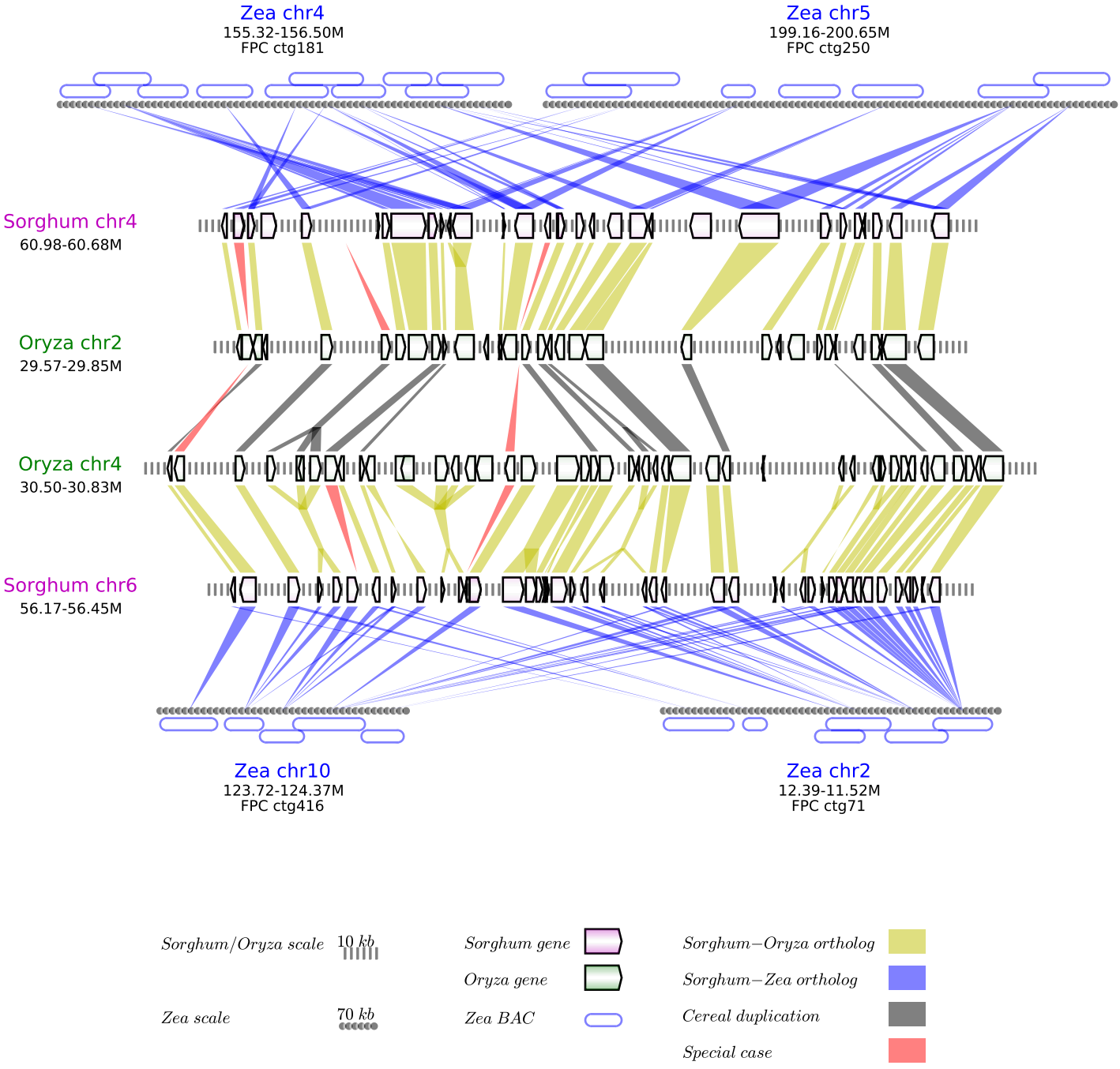
16 Project International Rice Genome Sequencing, *Nature* **436** (7052), 793 (2005).
17 G. Haberer, S. Young, A. K. Bharti et al., *Plant Physiology* **139**, 1612 (2005).
18 R. Bruggmann, A. K. Bharti, H. Gundlach et al., *Genome Research* **16** (10), 1241 (2006).
19 S. W. Roy and D. Penny, *Molecular Biology and Evolution* **24** (1), 171 (2007).
20 C. Lu, D. H. Jeong, K. Kulkarni et al., *Proceedings of the National Academy of Sciences of the United States of America* **105** (12), 4951 (2008).
21 R. Song, V. Llaca, and J. Messing, *Genome Research* **12** (10), 1549 (2002).
22 B. C. Meyers, A. Kozik, A. Griego et al., *Plant Cell* **15** (7), 1683 (2003).
23 D. Leister, *Trends in Genetics* **20** (3), 116 (2004).
24 N. C. Carpita and D. M. Gibeaut, *Plant Journal* **3** (1), 1 (1993); M. C. McCann and K. Roberts, in *The Cytoskeletal Basis of Plant Growth and Form*, edited by C. W. Lloyd (Academic Press, New York, 1991), pp. 109.
25 N. C. Carpita, *Annual Review of Plant Physiology and Plant Molecular Biology* **47**, 445 (1996).
26 S. P. Hazen, R. M. Hawley, G. L. Davis et al., *Plant Physiology* **132** (1), 263 (2003).
27 B. T. Zhao, R. Q. Liang, L. F. Ge et al., *Biochemical and Biophysical Research Communications* **354** (2), 585 (2007).
28 D. E. Nelson, P. P. Repetti, T. R. Adams et al., *Proceedings of the National Academy of Sciences of the United States of America* **104**, 16450 (2007).
29 A.H. Paterson, J.E. Bowers, and B. A. Chapman, *Proceedings of the National Academy of Sciences of the United States of America* **101**, 9903 (2004).
30 A. H. Paterson, B. A. Chapman, J. Kissinger et al., *Trends in Genetics* **22**, 597 (2006).
31 X. Wang, X. Shi, B. Hao et al., *New Phytologist* **165** (3), 937 (2005); The Rice Chromosomes 11 and 12 Sequencing Consortia, *BMC Biology* **3**, 20 (2005); Jun Yu, Jun Wang, Wei Lin et al., *PLoS Biology* **3** (2), e38 (2005).
32 X. Wang, H. Tang, J. E. Bowers et al., *Genetics* **177**, 1753 (2007).
33 N. K. Singh, V. Dalal, K. Batra et al., *Funct Integr Genomics* **7** (1), 17 (2007); Srinivasachary, M. M. Dida, M. D. Gale et al., *Theor Appl Genet* **115** (4), 489 (2007).
34 T. Tanaka, B. A. Antonio, S. Kikuchi et al., *Nucleic Acids Research* **36**, D1028 (2008).
35 S. Ouyang, W. Zhu, J. Hamilton et al., *Nucleic Acids Research* **35**, D883 (2007).
36 N. Huo, G. R. Lazo, J. P. Vogel et al., *Functional and Integrated Genomics* **8**, 135 (2007).
37 E. H. Margulies, J. P. Vinson, W. Miller et al., *Proceedings of the National Academy of Sciences of the United States of America* **102** (13), 4795 (2005); S. R. Eddy, *Plos Biology* **3** (1), 95 (2005).
38 F. Wei, E. Coe, W. Nelson et al., *PLoS Genet* **3** (7), e123 (2007).
39 N. Jannoo, L. Grivet, N. Chantret et al., *The Plant Journal* **50**, 574 (2007).
40 R. Ming, P. H. Moore, K. K. Wu et al., *Plant Breeding Reviews* **27**, 15 (2005).
41 H. C. Lohithaswa, F. A. Feltus, H. P. Singh et al., *Theoretical and Applied Genetics* **115**, 237 (2007).
42 A. H. Paterson, Y. R. Lin, Z. K. Li et al., *Science* **269** (5231), 1714 (1995).
43 Anonymous, 1997; Anonymous, 2002.

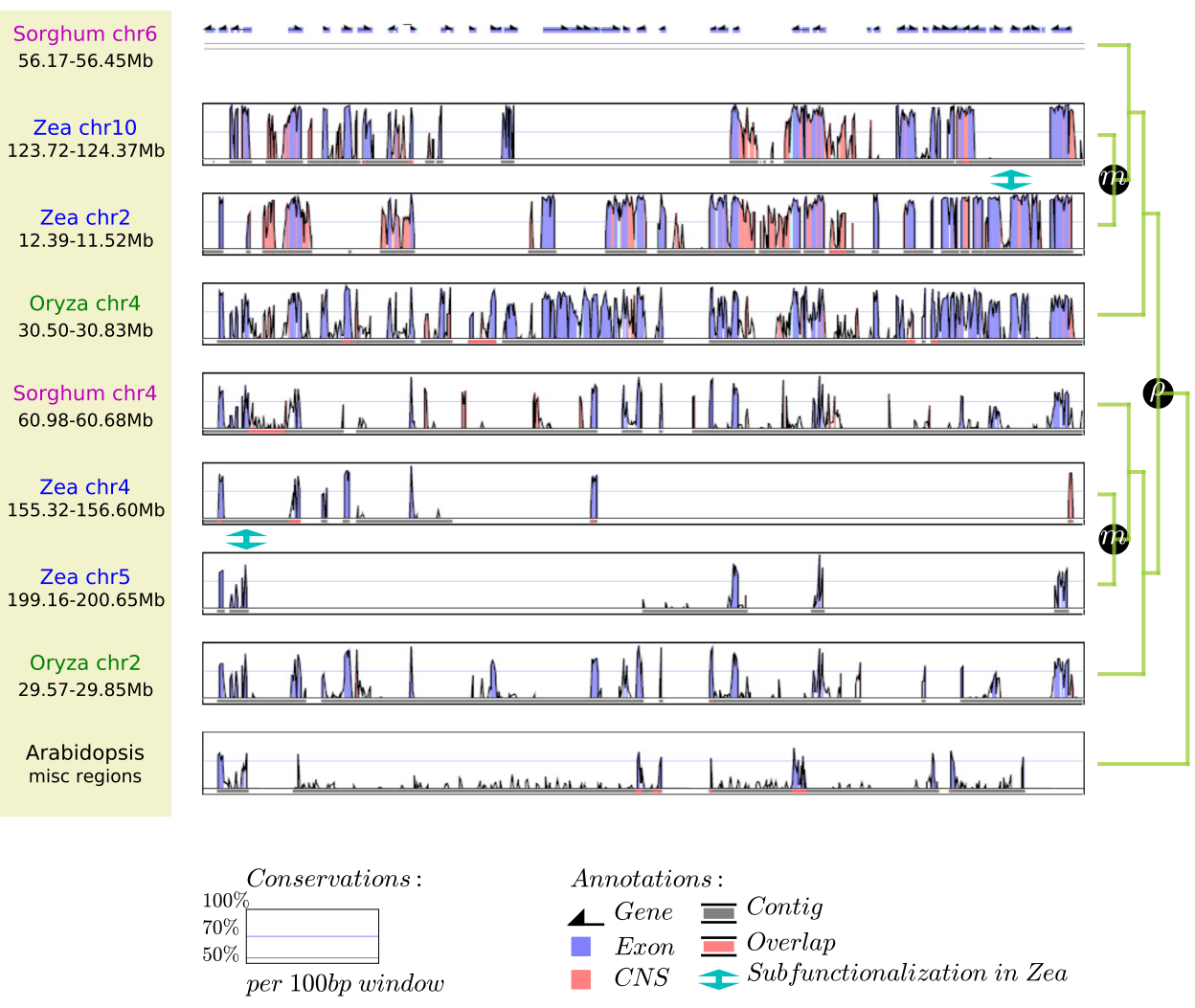

Saccharinae
Sorghum bicolor
Sorghum halepense
Sorghum propinquum
Saccharum (sugarcane)
Miscanthus
Zea (maize)
Panicum (switchgrass)
Genome duplication
ρ
Oryza (rice)
Brachypodium
Hordeum (barley)
Triticum (wheat)
Festuca (fescue)
Ananas (pineapple)

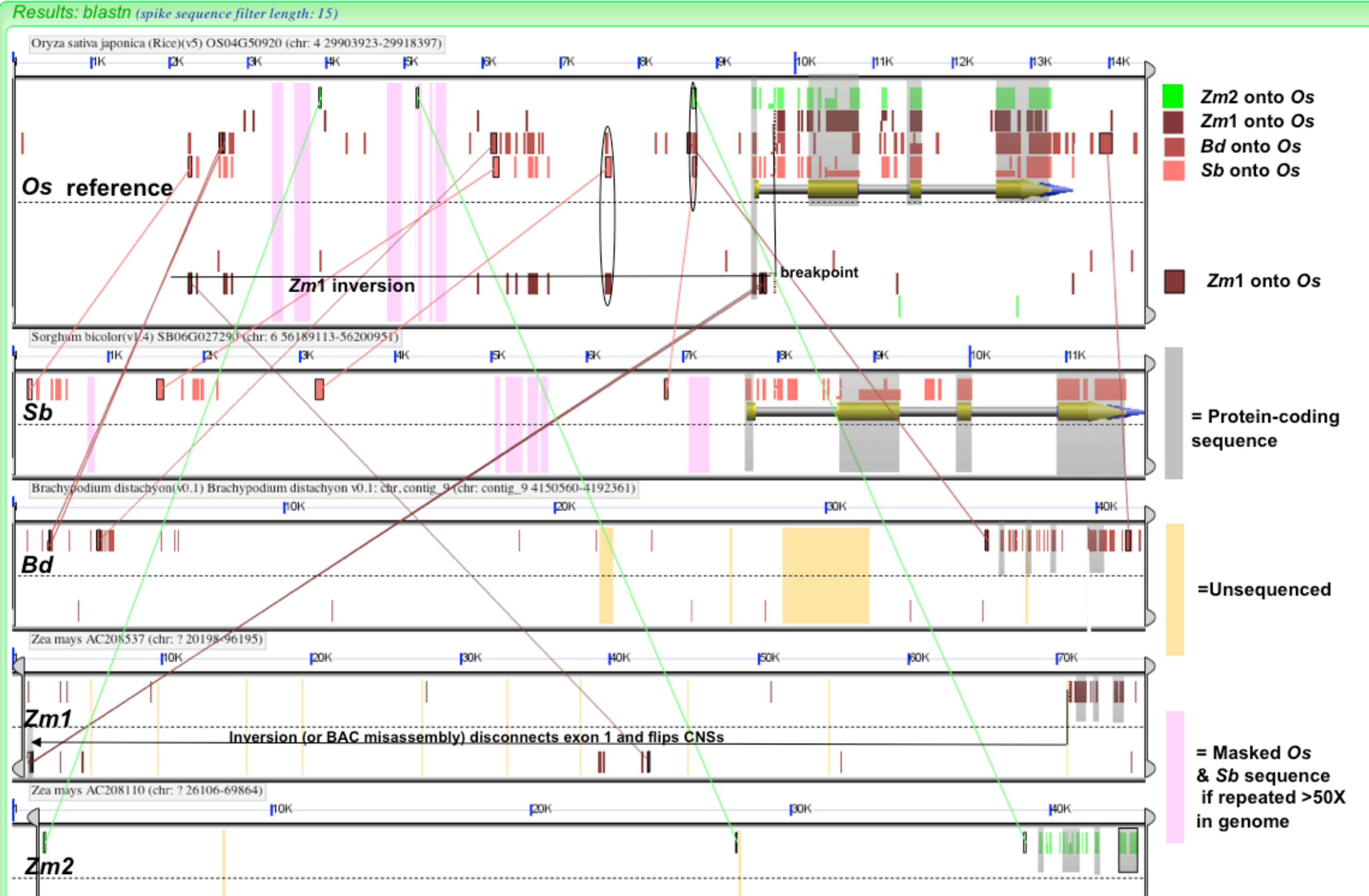

Chr 3
Cen38
Retrotransposons
DNA-Transposons
Genes (introns)
Genes (exons)
young LTR-RTs
full length LTR-RTs
LTR-RT/gypsy
LTR-RT/copia
DNA-TE/CACTA
CpG islands
DNA-TE/MITE
Genes (exons)
Paralogs
Paralogs
Genes (exons)
DNA-TEs/MITE
CpG islands
DNA-TE/CACTA
LTR-RT/copia
LTR-RT/gypsy
full length LTR-RTs
young LTR-RTs
Chr 9
0
20
40
60
[Mb]

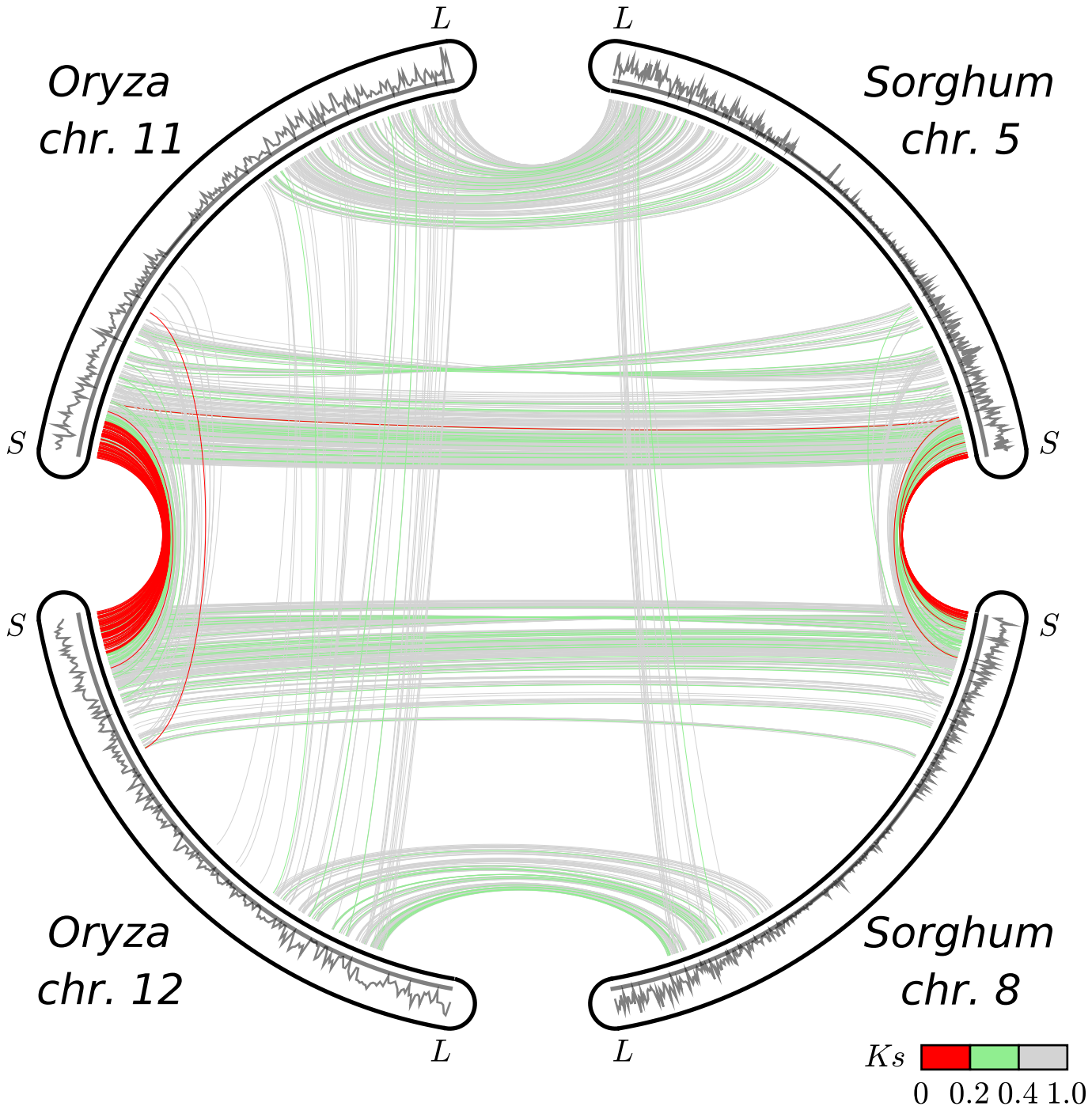

autonomous G118 "Mother" element
Transposase
ORF2
CACTA_G118_102, 9043 bp
carbonic anhydrase chloroplast
metal transporter Nramp6
CACTA_G118_101, 6698 bp
HP
CACTA_G118_104, 3609 bp
conserved HP
plant-specific domain
importin beta-1 subunit
CACTA_G118_110, 5088 bp
replication factor A
CACTA_G118_106, 6537 bp
replication factor A
CACTA_G118_111, 6491 bp
replication factor A
CACTA_G118_112, 6482 bp
carbonic anhydrase, chloroplast
HP
40S ribosomal S15,
CACTA_G118_114, 9829 bp
importin beta-1 subunit
CACTA_G118_116, 4244 bp
foreign gene fragments
> 30% identical
> 40% identical
> 50% identical
> 60% identical
with rice protein
Conserved TIR region
Non-autonomous derivatives
Other non-autonomous CACTAs
RNA polymerase Rpb1
CACTA_4965, 5530 bp
Os02g05880
HP
CACTA_6488, 6595 bp
2 kb


Arabidopsis
13,144
22,813
Sorghum
16,378 clusters
28,375 genes
Poplar
15,288
34,783
Rice
15,148
20,109
879
1153
49
1686
3983
634
229
2403
542
196
9503
25
631
139
96


Zea chr4
155.32-156.50M
FPC ctg181
Zea chr5
199.16-200.65M
FPC ctg250
Sorghum chr4
60.98-60.68M
Oryza chr2
29.57-29.85M
Oryza chr4
30.50-30.83M
Sorghum chr6
56.17-56.45M
Zea chr10
123.72-124.37M
FPC ctg416
Zea chr2
12.39-11.52M
FPC ctg71
Sorghum/Oryza scale
10 kb
Zea scale
70 kb
Sorghum gene
Oryza gene
Zea BAC
Sorghum−Oryza ortholog
Sorghum−Zea ortholog
Cereal duplication
Special case


Sorghum chr6
56.17-56.45Mb
Zea chr10
123.72-124.37Mb
Zea chr2
12.39-11.52Mb
Oryza chr4
30.50-30.83Mb
Sorghum chr4
60.98-60.68Mb
Zea chr4
155.32-156.60Mb
Zea chr5
199.16-200.65Mb
Oryza chr2
29.57-29.85Mb
Arabidopsis
misc regions
m
m
p
Conservations :
100%
70%
50%
per 100bp window
Annotations :
Gene
Exon
CNS
Contig
Overlap
Subfunctionalization in Zea


Results: blastn (spike sequence filter length: 15)
Oryza sativa japonica (Rice)(v5) OS04G50920 (chr: 4 29903923-29918397)
Os reference
Zm1 inversion
breakpoint
Zm2 onto Os
Zm1 onto Os
Bd onto Os
Sb onto Os
Zm1 onto Os
Sorghum bicolor(v1.4) SB06G027290 (chr: 6 56189113-56200951)
Sb
= Protein-coding sequence
Brachypodium distachyon(v0.1) Brachypodium distachyon v0.1: chr_contig_9 (chr: contig_9 4150560-4192361)
Bd
=Unsequenced
Zea mays AC208537 (chr: ? 20198-96195)
Zm1
Inversion (or BAC misassembly) disconnects exon 1 and flips CNSs
= Masked Os & Sb sequence if repeated >50X in genome
Zea mays AC208110 (chr: ? 26106-69864)
Zm2


Oryza
chr. 11
Sorghum
chr. 5
Oryza
chr. 12
Sorghum
chr. 8
L
L
S
S
S
S
L
L
Ks
0
0.2
0.4
1.0

Methods Summary

Genome sequencing and assembly. Approximately 8.5-fold redundant paired-end shotgun sequencing was performed using standard Sanger methodologies from small (~2-3 kb), and medium (5-8 kb) insert plasmid libraries, one fosmid library (~35 kb inserts), and two BAC libraries (insert size 90 and 108 kb). (Supplementary Note 1.)

Integration of shotgun assembly with genetic and physical maps. The largest 201 scaffolds, all larger than 39 kbp, excluding "N"s, and collectively representing 678,902,941 bp or 97.3% of all nucleotides, were checked for possible chimeras based on 5 independent lines of evidence, namely the sorghum genetic map, physical map, abrupt changes in gene or repeat family density, rice gene order, and coverage by BAC or fosmid clones, as detailed in Supplemental Note 2.

Repeat analysis. *De novo* searches were performed for LTR retrotransposons using LTR_STRUCT (pmid 12584121). *De novo* detection of CACTA-DNA transposons and MITEs used custom programs (Supplemental Note 3). Known repeats were identified by RepeatMasker (Open-3-1-8) (www.repeatmasker.org) with mips-REdat_6.2_Poaceae, a customized compilation of grass repeats that contained the new sorghum-specific LTR retrotransposons (mips.gsf.de/proj/plant/webapp/recat/). The insertion age of full length LTR-retrotransposons was determined from the evolutionary distance between 5' and 3' soloLTR derived from a ClustalW alignment of the two soloLTRs.

Protein-coding gene annotation. Putative protein-coding loci were identified based on BLAST [1] alignments of rice and Arabidopsis peptides and expressed sequence tags (ESTs) from sorghum and maize. The homology-based gene finder GenomeScan [2] was applied using maize-specific parameters. Predicted coding structures were merged with EST data from maize and sorghum using PASA [3].

Inter- and intra-genomic alignments. Comparative dot plots used ColinearScan [4] and multi-alignments used MCScan [5], applied to RAP2[33] (mapped representative models, 29389 loci) and *Sorghum bicolor* sbi1.4 annotation set (34496 loci). Pairwise BLASTP (E < 1e-5, top five hits), both within each genome and between the two genomes was used to retrieve potential anchors. *Zea* BAC sequences and FPC contig coordinates were downloaded from the Maize Genome Browser (http://www.maizesequence.org, release Jan. 7, 2008). *Sorghum* coding sequences were searched against *Zea* BACs for potential orthologous *Zea* genes using translated BLAT [6] with minimum score 100.

Supplemental Materials

Methods Summary..... pg. 1

Figure S1. Evolutionary context of sorghum and distinguishing features of the Saccharinae..... pg. 5

Supplemental Note 1. Genome sequencing data..... pg. 6
- **S1.1 DNA source and material preparation**
- **S1.2 shotgun library preparation (plasmid and fosmid)**
- **S1.3 BAC library sequencing**
- **Table S1. Shotgun sequencing statistics summary**

Supplemental Note 2. Genome assembly and map integration..... pg. 7
- **S2.1 Arachne assembly of whole genome shotgun dataset**
- **Table S2. Final summary statistics, map-integrated Arachne2 assembly**
- **S2.2 Manual curation of assembly and integration of map data**
- **Table S3: Sequence scaffold breaks made based on comparisons with physical map**
- **Table S4: Scaffold joins to reconstruct chromosomes, based on physical and genetic map**
- **S2.3 Telomeres**
- **S2.4 Completeness of assembly**
- **S2.5 Accuracy of the assembly in genic and repetitive regions**
- **Table S5. Comparison of the WGS assembly to randomly chosen BAC clones**
- **S2.6 Reconciliation of the assembly with genetic and physical maps, stress-testing based on synteny, and chromosome identification.**
- **Figure S2: Example of WGS assembly verification on one scaffold, and assembly of scaffolds into chromosomes.**
- **Table S6: Evidence for each scaffold join in chromosome assembly**
- **Table S7: Genome size evolution and distribution of recombination in sorghum, rice, and maize**
- **S2.7 Organellar sequences**
- **Table S8: Length and distribution of chloroplast DNA insertions on the Sorghum chromosomes**
- **Table S9: Length and distribution of mitochondrial DNA insertions on the Sorghum chromosomes**
- **S2.8 CpG Island Detection**

Supplemental Note 3. Repeat identification and characterization..... pg. 24
- **S3.1 Identification of LTR-retrotransposons**
- **Figure S3: Timing of LTR-Retrotransposon Insertions**

S3.2 MITEs
S3.2 Masking based on known repetitive sequences
S3.3 Masking based on over-represented 16-mers
Table S10: Repeat composition and major components of the sorghum genome in comparison to rice and maize
Table S11. Repeat composition by type
Table S12. Lineage specificity of transposons
Table S13. Repetitive content per chromosome
S3.4 CACTA Search Strategy
Figure S4. Dotplot of a small non-autonomous CACTA element sequence compared with itself
S3.5 Helitron identification
Table S14: Helitrons in sorghum and maize
S3.6 Tandem repeats
Table S15: Tandem repeats
S3.7 Repeat annotation and data integration
Figure S5: Genomic landscape of sorghum

Supplemental Note 4. Gene annotation and analysis..... pg. 35
S4.1 Structural gene calls in the Sorghum genome
S4.2 Gene identifiers
S4.3. Tandem gene clusters in sorghum
S4.4 Sorghum miRNA gene annotation
Table S16: miRNAs present in the Sorghum genome
Table S17: Position of known Sorghum miRNAs in the genome
Table S18: Position of newly detected miRNAs (paralog mapping) in the Sorghum genome
S4.5 Rice annotations for comparative analysis
S4.5.1 filtering of RAP2 rice annotation data
S4.5.2 TIGR5 gene set for rice
S4.6 Protein domains in the Sorghum genome
Table S19: Over- and underrepresented PFAM domains in the genome of *Sorghum bicolor*
S4.7 Protein family comparison across angiosperms
S4.8 Sorghum specific protein families
Table S20: Over- and underrepresented PFAM domains of *Sorghum* specific protein families

Supplemental Note 5. Gene structure and comparison with rice..... pg. 48

S5.1 Coding exon length distributions for sorghum (red) and rice (green)
Figure S6. Coding exon length distributions for sorghum (red) and rice (green)
Table S20: Statistics of sorghum gene composition

Figure S7. Nucleotide identity between sorghum transcripts and sugarcane (blue), maize (green), and rice (red) (based on assembled ESTs)
S5.3 Nucleotide identity between sorghum transcripts and sugarcane (blue), maize (green), and rice (red) (based on assembled ESTs)
S5.4 CISP identification

Supplemental Note 6. Identification and characterization of segments of conserved synteny..... pg. 50
Figure S8. Global dot-plot of *Oryza - Sorghum* and *Sorghum – Sorghum* genome alignments
S6.1 Pfam domains enriched in singleton or syntenic duplicated genes of sorghum

Supplemental Note 7. Timing and characterization of grass-specific genome duplication..... pg. 53

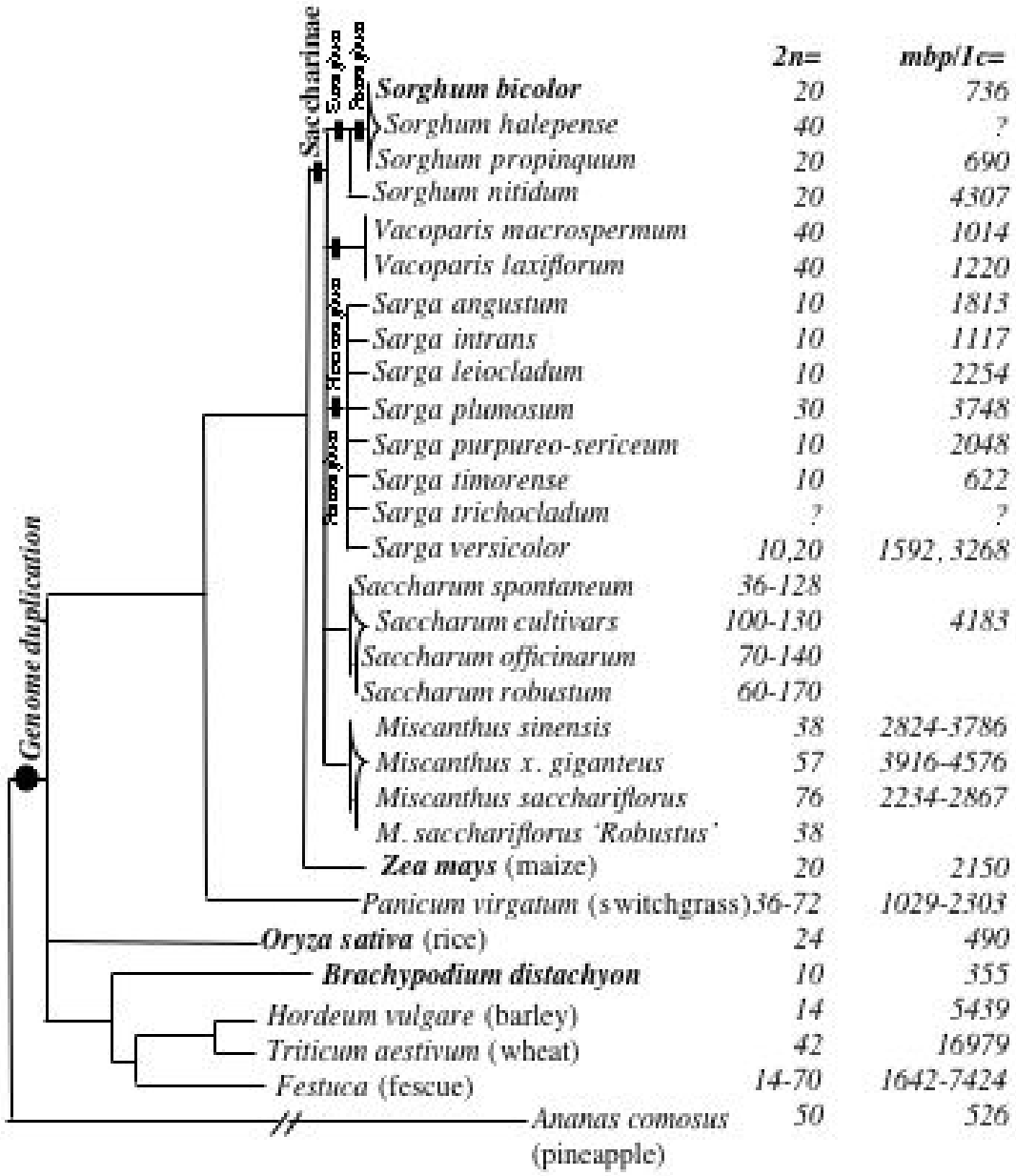
Supplemental Note 8. DNA alignments..... pg. 54
Table S21: Coverage of different interval of the sorghum genome by the alignments with the rice v.5.0 genome and 1.55Gbp of Maize BACs
Figure S9: Multiple VISTA conservation tracks among syntenic regions of plants
Figure S10: Grass conserved noncoding sequences (CNSs) are often far removed from the genes with which they associate

Supplemental Note 9. Evolution of C4 photosynthesis genes..... pg. 57
Table S22a. C4 genes identified in the sorghum genome
Table S22b. Sorghum C4 genes and their isoforms and their corresponding rice orthologs
Figure S11: Phylogeny of photosynthesis enzyme genes and their isoforms in sorghum, rice and maize

Supplemental Note 10. Evolution of cell wall synthesis genes..... pg. 62
Table S23. Comparative cell wall gene families of Arabidopsis, rice, sorghum, and maize
Figure S12. Cellulose synthase superfamily dendrogram for Arabidopsis, rice and sorghum.
Figure S13. GT31 dendrogram for Arabidopsis, rice, and sorghum.

Supplemental Note 11. Sorghum-sugarcane microcolinearity. ..... pg. 66
Figure S14. Collinear alignment between sugarcane BAC sequences and their sorghum counterparts
Figure S15. Tandem gene duplication in the sugarcane or sorghum genome

Figure S1. **Evolutionary context of sorghum and distinguishing features of the Saccharinae.** Branch lengths above the species level were computed by aligning EST assemblies from the TIGR PlantTA collection (plantta.tigr.org), and estimating the transversion rate at fourfold synonymous sites using a Jukes-Cantor correction for multiple transversions, and creating a phylogenetic tree with the neighbor-joining method as implemented in Phylip (evolution.genetics.washington.edu/phylip.html). Phylogenetic interpretation is from [7]; the ranges of genome size estimates are from [8] and direct measurements by flow cytometry *(S. propinquum, Miscanthus spp.).*



Supplemental Note 1. Genome sequencing details

We sequenced the genotype BTx623, a largely-homozygous breeding line released by Texas A&M University [9], which figures in the pedigrees of many elite sorghum genotypes and has been widely used in sorghum genomics research.

S1.1 DNA source and material preparation

Nuclear DNA was isolated from sorghum BTx623 seedlings as described [10] with minor modifications (see http://www.mgel.msstate.edu/pdf/nucl_dna.pdf). To remove potential carbohydrate contaminants that may have precipitated with the DNA, the isolated nuclear DNA was dissolved in 0.03 M sodium phosphate buffer (SPB) and loaded onto a hydroxyapatite column equilibrated with 0.03 M SPB. The DNA was washed with 10 column volumes of 0.03 M SPB and 10 column volumes of 0.12 M SPB, and then the DNA was eluted from the column by the addition of 0.5 M SPB [11] for procedures associated with hydroxyapatite chromatography). The nuclear DNA was transferred into 10 mM Tris buffer (pH 8.0) using a Centriplus YM-30 column (Millipore, Billerica, MA USA).

S1.2 shotgun library preparation and sequencing (plasmid and fosmid)

Plasmid- and fosmid-end sequencing was performed using standard library protocols and Sanger dye-terminator chemistries on the ABI-3730 and MegaBACE 4000 sequencing instruments. Sequencing totals are shown in Table S1, with insert sizes estimated self-consistently from the shotgun assembly using Arachne2 [12]. Sequence coverage is computed by counting the phred20 bases for each aligned trace and dividing that number by the assembled consensus bases. This avoids a priori estimates of the genome size, although collapsed repeats can still lead to an overestimate of coverage. In this calculation, scaffolds that fall outside of the normal coverage range are ignored. High quality (HQ) reads are longer than 200 bp free of vector sequence and with PhredQ>20. All traces for this project were deposited in the NCBI Trace Archive.

S1.3 BAC libraries and sequencing

A ~11x BAC library for BTx623 was prepared and fingerprinted previously, also hybridizing overgo probes from most genetically-mapped sequence tagged sites to the BACs to align the physical and genetic maps [13]. Paired-end sequences for total of 96,870 BAC clones (spanning 10.31 Gb ~ 13.5X clone coverage) were generated herein using standard chemistries on ABI-3730 sequencing instruments. All traces for this project were deposited in the NCBI Trace Archive.

Table S1. Shotgun sequencing summary statistics

Library Type	Insert Size (kbp)	Total Reads	Sequence Coverage	Reads <200 Phred 20s	All Vector Reads	Unpaired HQ Reads	Paired HQ Reads
Small insert	2.44+/-0.39	4,817,407	3.74x	363,104	102,781	236,912	4,114,610
Medium Insert I	6.40+/- 0.53	2,661,374	2.32x	137,325	82,071	67,246	2,374,732
Medium Insert II	6,881+/-0.59	2,149,803	1.72x	179,628	68,125	89,976	1,812,074
Medium Insert III	8.61+/-0.76	18,144	0.01x	1,048	293	695	16,108
Fosmid	34.7+/-3.8	850,443	0.52x	172,321	8,826	63,234	606,062
BAC/SB_BBc	108.0+/-21.8	193,920	0.17x	8,791	3,563	4,822	176,744
BAC/SB_BBd	91.0+/-25.0	26,112	0.02x	3,613	4,182	1,321	16,996
Total		**10,717,203**	**8.50x**	**865,830**	**269,841**	**464,206**	**9,117,326**

S2. Genome assembly and map integration

S2.1 Arachne assembly of whole genome shotgun dataset

An initial whole genome shotgun (WGS) assembly was built with Arachne2 [12] v.20060705 with 48-mers that occurred 65 or more times in the dataset considered repetitive; no error correction; and the option to remove and replace reads deemed wholly repetitive. We made modifications to the repeat identification process to allow Arachne2 to better identify and correct misassembled repetitive elements.

The resulting assembly is denoted Sbi1 and is deposited in Genbank as accession number ABXC00000000, and can also be obtained at www.phytozome.net/sorghum. A preliminary assembly using only a partial dataset was made available at phytozome in January 2007. This “Sbi0” assembly was transient and is superceded by Sbi1. All analyses in this manuscript refer to the Sbi1 assembly.

Table S2. Final summary statistics, map-integrated Arachne2 assembly.
Main genome contig total: 12,873
Main genome contig N/L50: 958 contigs longer than 195.4 KB
Main genome contig sequence total: 697.6 MB

Main genome scaffold total: 3,304
Main genome scaffold N/L50: 6 scaffolds longer than 62.4 MB
Main genome scaffold sequence total: 738.5 MB
These scaffolds include 24.2 MB of centromere spacers.
The estimated gap % without centromere spacers is 2.35%.

After breaking at the 28 points of incongruity with the physical map (see below), the contig N50 is 1,013 and L50 is 187.1 kb; the scaffold N50 is 35 and L50 is 7.0 Mbp.

Minimum Scaffold Length	Number of Scaffolds	Number of Contigs	Total Sequence Size*	Total Non-Gap Bases	%Scaffold Size in Non-Gaps
All	3,304	12,873	738,540,932	697,579,688	94.45%
1 kb	3,304	12,873	738,540,932	697,579,688	94.45%
2.5 kb	3,070	12,602	738,070,256	697,126,337	94.45%
5 kb	1,247	10,294	731,651,470	690,813,104	94.42%
10 kb	604	9,334	727,157,266	686,514,747	94.41%
25 kb	170	8,620	720,919,732	680,464,262	94.39%
50 kb	122	8,475	719,160,869	679,159,609	94.44%
100 kb	83	8,281	716,381,796	677,191,673	94.53%
250 kb	47	7,947	710,617,843	672,901,797	94.69%
500 kb	32	7,694	705,483,650	669,211,941	94.86%
1 Mb	19	7,356	696,834,577	662,602,196	95.09%
2.5 Mb	15	7,208	690,721,313	656,841,816	95.10%
5 Mb	13	7,084	682,507,325	648,764,287	95.00%

* Includes estimated centromere gap bases.

S2.2 Manual curation of assembly and integration of map data

After the Arachne assembly, 28 breaks (See S2.4 below) were made and 117 manual joins were made. These include ten gaps inserted for unassembled centromeres based on genetic and physical map data. The size of the centromere was estimated for each chromosome based upon the amount of centromeric sequence already assembled for that chromosome. The main genome is in 10 chromosomes along with ~3,000 small unmapped pieces, totaling 697.6 Mb. The unmapped sequences contain fewer than ~200 bona fide protein coding genes with homology to rice genes.

Table S3: Sequence scaffold breaks made based on comparisons with physical map.

Scaffold number	End of first segment	Start of second segment
1	6,773,471	6,774,471
1	16,715,462	16,716,462
1	17,049,685	17,050,685
2	6,769,918	6,770,918
3	14,348,828	14,349,828
5	6,309,929	6,310,929
7	1,932,250	1,933,250
7	10,109,836	10,110,836
8	5,100,008	5,101,008
8	12,471,542	12,472,542
11	930,223	931,223
12	11,445,401	11,446,401
15	6,665,054	6,666,054
15	8,885,477	8,886,477
21	1,527,622	1,528,622
23	3,601,417	3,602,417
24	132,868	133,868
26	398,727	399,727
28	7,342,067	7,343,067
31	1,632,238	1,633,238
32	6,399,433	6,400,433
34	4,368,763	4,369,763
46	3,760,727	3,761,727
47	99,836	100,836
49	2,737,864	2,738,864
58	4,088,464	4,089,464
65	2,903,582	2,904,582
91	340,379	341,379

Table S4: Scaffold joins to reconstruct chromosomes, based on physical and genetic map. Scaffold numbers represent original Arachne assembly after manual breaking. When included, decimals refer to sub-scaffolds after breaking. F and R indicate forward and reverse, respectively. The ten chromosomes are named by their chromosome numbers [14], reconciled with the leading sorghum genetic maps [15].

Chromosome number	p-arm	Centromere gap inserted as N's	q-arm
1	57R 1.4R 7.2R 95R 52F 26.1R 32.1F 94R	4,800,000	101F 21.2F 67F 73F 105F 49.1R 58.1R 97R 89F 78F 24.1R 47.2R 100F
2	40R 23.2R 46.1R 5.1F 13R	100,000	77F 11.1R 54F 61F 53F 11.2F 1.1R 37R
3	30R 18R 34.2R 2.2F	300,000	0R 7.1R 34.1F 2.1R
4	16F 42R 99F 5.2F	4,300,000	4F 72F 60R 43R 59F
5	10F 27F 69R 24.2R 48R	2,200,000	25F 49.2F 26.2F 90F
6	85F 14F 32.2R 36F	100,000	22F 23.1F 46.2F 15.1F 51R 118F 50F 62R 74R
7	103R 45R 92R 15.2R 8.3R 28.1F 8.2F 8.1F	3,600,000	9F 80F 15.3F 106R 125R 87R 56R 144R 1.3F 7.3F 123F
8	76R 79F 63F 19F	2,500,000	41R 3.1F 104F 21.1F 33F 83F
9	81F 88F 71F 12.1F 31.1R 12.2R 31.2F	3,200,000	115R 65.1R 44F 1.2R 84F 3.2F
10	20R 126R 65.2R 70F 75F 17F 86F	3,100,000	6F 68R 39R 119R 64R 29R 91.2R

S2.3 Telomeres. The sorghum telomere signature sequence is $(AAACCCT)_N$.

Chromosomes 1, 4, 5, 7,10 show evidence of having both telomeres attached; chromosomes 2, 3, 6, 8, and 9 include only one telomere in the assembly.

Chromosome	P telomere	Q telomere
1	Yes	Yes
2	No	Yes
3	Yes	No
4	Yes	Yes
5	Yes	Yes
6	No	Yes
7	Yes	Yes
8	No	Yes
9	Yes	No
10	Yes	Yes

S2.4 Completeness of assembly

To assess the completeness of the *S. bicolor* assembly, we aligned 20,417 *S. bicolor* transcript assemblies from the TIGR PlantTA gene indices using BLAT [16] against the 16-mer-repeat-masked sequence. Only 911, or 4.4%, did not map to the genome assembly.

Of these, 756 have a hit to Uniprot90. Only 51 were shown to have any similarity to known plant sequences, with the remainder dominated by hits to fungal genes (related to the genus *Fusarium*) or other likely contaminants of available sorghum cDNA libraries.

No hits to UniProt	155
Fungal	517
Lower eukaryote	89
Animal	76
Plant	51
Bacteria	15
Algal	7
Viral	1
Total TIGR Sorghum transcript assemblies that do not hit genome assembly	911

If we assume that the 51 hits to plant genes, 7 algal hits, and the 155 PlantTA's that don't hit UniProt90 together represent an overestimate of the missing protein-coding loci in the sorghum genome, then we have missed only at most ~1%.

S2.5 Accuracy of the assembly in genic and repetitive regions

To evaluate the accuracy of the assembly on a local scale, 31 BAC clones were subcloned into ~3kb insert plasmid clones and end-sequenced using ABI3730 Sanger methods, and finished to Bermuda standards by primer walking and gap closure.

Comparison of the assembly to these randomly chosen BAC clones showed excellent coverage and sequence-level accuracy (Table S5). 98.46% of the bases were represented in the assembly exactly as they appeared in the clones. When we exclude gap-adjacent, AT string, and marked low quality sequence the error rate is lower than 1 in 10,000 bp. However, the area covered by the finished clones includes 4 assembly collapses on repetitive elements which account for 35,040 of the non-matching bps in the 3.3 Mb surveyed (~1%) and one finished clone deletion of 4,223 bps.

Nearly 2/3 of the "missing" region from clone 4002310 can be found scattered throughout the genome, and represents repetitive regions that were not accurately captured in the whole genome assembly.

Clones 4000659 and 4000660 were the same clone accidentally sequenced twice; the only difference in these finished clones is the length of an (AT)n microsatellite.

Table S5. Comparison of the WGS assembly to randomly-chosen BAC clones.

PID	SIZE	START	STOP	DIR	CHR	START	STOP	IDENT	MISS	ERROR	GAP	EXTRA	Accuracy
>4000658.2	127,914	0	127,914	-	1	9,739,369	9,866,665	127,277	637	14	0	5	99.99
>4002334.3	128,769	0	128,769	-	1	12,137,156	12,265,625	128,156	613	8	603	82	99.76
>4002335.2	120,503	0	120,503	+	1	12,352,549	12,474,452	120,325	178	29	140	60	99.85
>4002313.1	128,409	0	128,409	-	1	12,447,249	12,575,655	128,406	3	0	0	0	100
>4002299.4	136,055	0	136,055	+	1	12,606,253	12,750,391	136,016	39	11	25	3	99.97
>4000656.3	129,695	0	129,695	+	1	16,983,689	17,113,180	129,132	563	9	341	9	99.72
>4000657.3	117,931	0	117,931	+	1	68,817,224	68,934,972	117,391	540	9	361	0	99.7
>4002300.2	114,848	0	114,848	+	3	46,423,819	46,534,299	110,327	4521	2	172	1	99.86
>4002303.1	106,227	0	106,227	+	3	46,498,351	46,604,427	105,498	729	26	291	202	99.46
>4002310.2	121,989	0	121,989	-	3	46,675,710	46,789,249	112,635	9354	8	405	299	99.2
>4002316.4	100,445	0	100,445	+	3	46,756,550	46,857,772	99,931	514	2	451	531	99.49
>4002445.5	112,839	0	112,839	-	3	61,149,622	61,267,708	111,929	910	94	640	4,223	99.19
>4002446.3	96,424	0	96,424	+	4	14,631,428	14,727,858	96,423	1	0	0	7	100
>4002328.7	64,213	0	64,213	+	4	14,731,102	14,795,300	64,194	19	4	0	0	99.99
>4000662.5	108,555	0	108,555	+	4	61,840,377	61,950,973	106,932	1623	125	1.386	16	98.5
>4000659.1	105,241	0	105,241	-	5	7,420,443	7,526,019	105,119	122	37	0	320	99.88
>4000660.2	105,215	0	105,215	+	5	7,420,443	7,526,019	105,119	96	37	0	320	99.91
>4000653.5	138,518	0	138,518	+	8	1,361,065	1,494,011	131,842	6676	38	990	5	99.17
>4000655.3	135,209	0	135,209	+	8	1,959,194	2,094,545	135,140	69	45	0	166	99.95
>4000661.2	137,889	80,919	119,588	+	8	43,520,874	44,268,271	38,190	479	450	0	664.829	98.76
>4000663.5	122,145	37	122,145	-	8	53,703,811	53,826,194	121,835	273	25	0	207	99.78
>4002308.25	117,078	0	117,078	-	8	53,886,081	54,002,976	116,754	324	30	152	11	99.88
>4000664.11	70,712	0	70,712	-	9	5,655,664	5,726,468	70,702	10	8	0	94	99.99
>4002337.14	145,423	0	145,423	-	9	20,978,313	21,117,184	138,793	6630	72	0	6	99.94
>4002317.3	149,983	0	98,979	-	9	21,035,950	21,396,759	98,830	149	130	0	261,849	99.85
>4002305.2	115,915	0	115,915	+	9	21,090,041	21,205,951	115,891	24	19	0	0	99.98
>4002441.4	128,738	0	123,196	+	9	21,183,013	21,298,372	115,337	7859	22	0	0	99.98
>4002450.1	112,916	0	112,916	-	9	21,430,150	21,543,374	112,916	0	0	0	208	100
>4002336.3	105,211	0	105,211	+	9	21,522,917	21,628,067	104,854	357	4	0	192	99.72
>4002309.3	123,072	0	123,072	-	9	21,662,660	21,788,680	123,047	25	23	0	1950	99.98
>4002301.3	130,057	0	130,057	+	9	21,788,674	21,918,332	127,851	2206	13	2190	4	98.61

S2.6 Reconciliation of the assembly with genetic and physical maps, stress-testing based on synteny, and chromosome identification.

The robustness of assembly of the 201 largest scaffolds (representing 678,902,941 bp or 97.3% of all nucleotides) was tested based on several independent lines of evidence.

Sequences from 2,050 genetically-mapped RFLP probes from a 2,512 locus map that defines 61.5% of the recombination events in the underlying population [17] were compared to the longest 201 scaffolds via BLAST (blastn E≤1E-6), plotting the corresponding locations for the top three hits for each sequence on a representation of the scaffold (for example, see one scaffold in Figure S2). Multiple hits were plotted for each RFLP probe sequence because some probes map to multiple loci, and other probes that only map to single loci have additional copies that were not polymorphic in the mapping population.

A physical map consisting of 1,869 contigs assembled from an 11x coverage BAC library by BAC fingerprinting and overgo hybridization [13] was compared to the assembly by superimposing paired-end sequences from the physically mapped BACs onto the 201 sequence scaffolds. A dot was plotted for each BAC end corresponding to its contig in the physical map and its position in the scaffold assembly. BAC ends were plotted in black, green, or red, respectively, for physical contigs that had 3 or more BAC ends going to 1, 2, or 3 or more different sequence scaffolds. Of the 1,869 physical contigs, only 122, 18, and 6 corresponded to two, three, and four different sequence scaffolds, respectively. This agreement between the physical contigs and the independently constructed sequence assembly strongly supported that each data type accurately represented the sorghum DNA. Incongruities such as a physical map contig mapping to the center of two different sequence scaffolds indicated loci where either the physical map contig or the sequence scaffold was assembled incorrectly. Cases where the end of a sequence scaffold occurred in the middle of a physical map contig could be used as a hint as to which scaffold it should be assembled with (in combination with support by other evidence). A total of 37 of the 117 joins were made in this manner.

The next line of evidence was to plot the gene density (as represented by the best hit to sorghum ESTs); matches to two abundant retroelements Candystripe1 (which corresponded strongly with gene rich regions - [18]), and Retrosor6 which corresponded with gene poor heterochromatin [11]; or to CEN38, a centromeric repeat [19]. In general the chromosome ends tended to be rich in Candystripe1 and ESTs, with interstitial levels gradually decreasing accompanied by progressively higher densities of Retrosor6, and finally with stretches of CEN38 in the centromeric regions. The transitions from gene rich to gene poor was generally gradual -- abrupt transitions from very low densities of Retrosor6 to very high densities provided further support of assembly errors that were already suspected due to other lines of evidence.

The rice genome was previously known to show good collinear synteny to sorghum with a limited number of macro-scale rearrangements [20]. Predicted rice genes (TIGR Version 4.0, excluding retrotransposon related genes) were plotted onto the sorghum scaffolds by BLAST (tblastx E≤1E-6). The best and second-best hits of a rice gene were plotted as black and red dots, respectively. In many regions stretches of best hits corresponding to the orthologous and second best hits corresponding to paralogous regions were evident, due to ancient polyploidy [20]. Synteny was used as an additional test of the sequence assembly, as a scaffold would not be expected to show long stretches of rice-sorghum synteny if incorrectly assembled. It must be emphasized that synteny information was

only used to support other lines of evidence, and the assembly never exclusively relied on synteny to support any genome arrangement.

A total of 109 regions were not spanned by any BAC clones and 59 were not spanned by fosmid clones (44 lacked both BACs and fosmids). These regions, assembled exclusively from the smaller clone libraries, would be extremely sensitive to misassembly associated with duplicated or repetitive DNA, and were noted as further support of suspected assembly errors.

Collectively, these independent lines of evidence identified 28 assembly errors, in all cases relying on multiple lines of evidence. After breaking the scaffolds at the 28 points that appeared to be incorrect joins, the resulting 229 scaffolds and scaffold pieces were assembled into chromosomes where possible based on the physical map, genetic map, rice synteny and genome structure (as represented by the gene and repeat distribution), inferring joins based on at least two independent lines of evidence. Finally in the process of development of the genome assembly 3 previous automated assemblies had been made with incomplete sets of the sequence data. In many cases the different assemblies showed different breakpoints, with a region being assembled in one assembly differently from another.

In total 127 of the scaffolds could be assembled into chromosomes representing 625,636,247 bp or 89.7% of all basepairs. Based on cytological evidence [14] the resulting assemblies were oriented to place the shorter chromosome arm at the top. In total 117 joins could be inferred between adjacent scaffolds based on the multiple lines of evidence discussed above, orienting all 127 scaffolds and providing an initial representation of the 10 sorghum chromosomes. Different pieces of evidence were used for each join as listed in Table S6. In general no more than one join per chromosome lacked support from two or more types of evidence. All such unsupported joins were at the centromere, resulting in two largely complete chromosome arms that could be assigned and oriented by genetic markers. Several large scaffolds could not be assembled into the chromosomes, the 5 largest being 8.8, 7.2, 7.1, 4.6, and 3.6 mbp respectively. Most likely two of these large unanchored scaffolds belong to chromosome 1, the only one that was notably smaller than the size predicted by cytology.

The remaining 102 scaffolds tended to be much smaller and were predominantly centromeric, with 85 containing major stretches of the centromeric repeat CEN38 [19]. Overall the chromosomes as assembled contain, respectively, approximately 0, 5.5, 4.5, 0.5, 2.6, 5.2, 1.2, 2.3, 1.6, and 1.7 mbp of centromeric repeats (based on the size of the region of dense Cen38 element abundance, with most Cen38 elements accounted for by these regions. This totaled 25.1 mbp of centromeric regions assembled into chromosomes. A total of 1362 unassembled sequence scaffolds (representing 34.94 mbp or 5% of all basepairs) are presumably centomeric based on the presence of Cen38 elements. Anchored and unanchored scaffolds together total about 60 MBP. However about 20% of this was gaps of “Ns” in the assembly, suggesting that the true total size of the centromeres is about 48 mbp. The centromeres of two chromosomes (2 and 6) are markedly larger than the average of 4.8 mbp, either due to false assembly or variation in centromere size. To account for missing centromeric DNA in chromosome assemblies,

gaps of Ns was incorporated into the assembly of each chromosome to bring the centromere size up to the average 4.8 mbp. For the 10 chromosomes, respectively, this added 4.8, 0.1, 0.3, 4.3, 2.2, 0.1, 3.6, 2.5, 3.2, and 3.1 mbp of Ns.

Alignment of the sorghum and rice sequence scaffolds, and the published maize physical map [21], to the respective genetic maps, permitted a comparative analysis of genome size evolution.

Figure S2 (next page after caption). Example of WGS assembly verification on one scaffold, and assembly of scaffolds into chromosomes.

Scaffold-7 from the WGS assembly is shown. The far left scale shows the location on the scaffold in MBP. The left hand side of the figure shows synteny to rice genes with the horizontal axis representing the rice genes in order over the 12 rice chromosomes. A black dot represents the location of a best hit of a rice gene to the sorghum sequence assembly and a red dot represents the second best hit. Linear patterns of dots represent segments of synteny to this scaffold. The lines predominantly composed of black dots correspond to rice-sorghum orthologous segments, and the lines predominantly consisting of red dots represent homoeologous (or paralogous) synteny resulting from the "rho" paleopolyploidy event common to the grasses [20]. Scattered dots tend to correspond to single gene duplications and translocations that add noise to the general pattern of synteny. Horizontal lines in this section represent portions of the scaffold that were not spanned by large insert clones, with red horizontal lines not spanned by BACs and blue horizontal lines not spanned by fosmid clones. Such areas not spanned by large insert clones are prone to mis-assembly by Arachne2 in the automatically generated WGS scaffold, due to repeats that are too long to be disambiguated by shorter-insert clones, such as recently duplicated copies of ~10-kb Retrosor-6 [22] or other retrotransposons that are nearly identical.

The middle section of the figure represents a reference genetic map of sorghum [17]. Each colored horizontal line in this section represents a marker on the genetic map. Line colors correspond to different sorghum chromosomes with color coding provided in a legend near the bottom of the figure, and the lengths of the lines corresponding to the genetic map position in centimorgans. Accordingly, different genetic markers that are closely linked on the genetic map would be expected to have lines of similar length. We note that since roughly 1/3 of the loci on the genetic map were from probes that mapped to multiple locations, some genetically mapped loci are paralogs and are expected to be inconsistent with the sequence assembly, as it is with rice synteny (Paterson et al 2004).

The next section represents the FPC based physical map of Sorghum [13]. In this case each dot represents a sequenced BAC end, with the horizontal position of the dots corresponding to the FPC contig number within the physical map. FPC contigs that had >4 BAC ends matching more than one sequence assembly scaffold are plotted with green dots while all others are plotted with black dots.

The far right portion of the figure shows distributions of Retrosor-6 repeats (green), Cen-38 repeats (red, not present in this scaffold), Candystripe-1 elements (light blue), sorghum ESTs (black) and sequence gaps in the scaffold (dark blue).

Scaffold-7 shows 3 abrupt changes in rice-sorghum synteny, corresponding to roughly 2 MBP, 4.4 MBP, and 10.2 MBP. The synteny breakpoints at 2 MBP and 10.2 MBP also correspond to regions of the scaffold not spanned by large insert clones (BACs, and fosmids and BACs respectively). The 2 MBP and 10.2 MBP synteny discontinuities also correspond to breakpoints in alignment to the sorghum genetic map, with the top portion of the scaffold corresponding to sorghum chromosome 3 above the 2 MBP breakpoint, to chromosome 1 after the first breakpoint, and finally to chromosome 7 after the 10.2 MBP breakpoint. Finally only 6 FPC contigs on this scaffold did not agree with the sequence scaffold assembly (in green), and of these 4 were clustered in pairs around the 2 MB and 10.2 Mb points not spanned by large insert clones –which are also synteny breakpoints. The synteny breakpoint around 4.5 MBP was not suggested by any other lines of evidence to be an error in the scaffold assembly, and was consistent with a previously-identified genomic rearrangement distinguishing rice and sorghum [17].

The various lines of evidence shown in the figure were used to break the automatically assembled scaffold-7 into 3 parts at 2 MBP and 10.2 MBP. Similar figures were examined for the 201 largest scaffolds to verify scaffold assemblies. The same lines of evidence were then used to reassemble the scaffolds and scaffold parts into chromosomal assemblies.

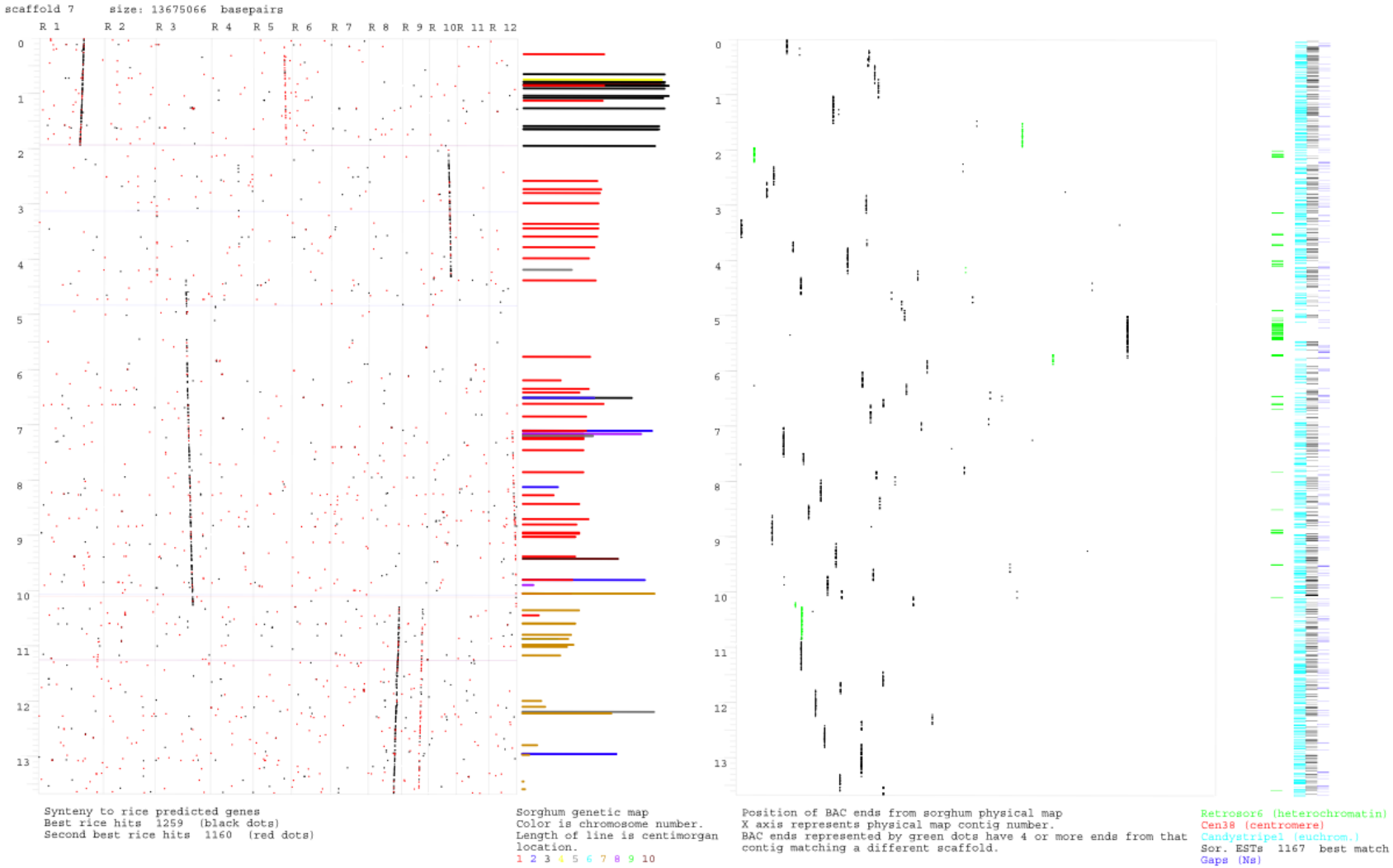

scaffold 7 size: 13675066 basepairs
R 1 R 2 R 3 R 4 R 5 R 6 R 7 R 8 R 9 R 10R 11 R 12
0 1 2 3 4 5 6 7 8 9 10 11 12 13
Synteny to rice predicted genes
Best rice hits 1259 (black dots)
Second best rice hits 1160 (red dots)
Sorghum genetic map
Color is chromosome number.
Length of line is centimorgan location.
1 2 3 4 5 6 7 8 9 10
Position of BAC ends from sorghum physical map
X axis represents physical map contig number.
BAC ends represented by green dots have 4 or more ends from that contig matching a different scaffold.
Retrosor6 (heterochromatin)
Cen38 (centromere)
Candystripe1 (euchrom.)
Sor. ESTs 1167 best match
Gaps (Ns)

Table S6: Evidence for each scaffold join in chromosome assembly

Sorghum chromosome	Scaffold numbers flanking inferred join	Affected contigs were joined in alternate (preliminary, or less stringent) sequence assemblies	Synteny to rice across gap	Genetic map shows close linkage between scaffolds	Gap is spanned by physical map contig	Note
1	57R-1.4R	No	Strong	Yes	No	Telomere
1	1.4R-7.2R	Yes	Strong	Yes	Yes	
1	7.2R-95R	Yes	Strong	Yes	Yes	
1	95R-52F	No	Strong	Yes	No	
1	52F-26.1R	Yes	Strong	No	No	
1	26.1R-32.1F	No	Strong	No	No	
1	32.1F-94R	Yes	Weak	No	Yes	
1	94R-101F	No	Weak	No	No	Centromere
1	101F-21.2F	Yes	Weak	Yes	Yes	
1	21.2F-67F	No	Weak	Yes	No	
1	67F-73F	No	Strong	Yes	No	
1	73F-105F	No	Strong	No	No	
1	105F-49.1R	Yes	Strong	No	Yes	
1	49.1R-58.1R	Yes	Strong	Yes	Yes	
1	58.1R-97R	No	Strong	Yes	No	
1	97R-89F	Yes	Strong	Yes	No	
1	89F-78F	Yes	Strong	Yes	No	
1	78F-24.1R	Yes	Strong	No	Yes	
1	24.1R-47.2R	Yes	Strong	No	Yes	
1	47.2R-100F	Yes	Strong	Yes	Yes	Telomere
2	37F-1.1F	No	Strong	Yes	No	
2	1.1F-11.2R	Yes	Strong	Yes	Yes	
2	11.2R-53R	Yes	Strong	Yes	Yes	
2	53R-61R	No	Weak	Yes	No	
2	61R-54R	Yes	Weak	Yes	Yes	
2	54R-11.1F	Yes	Weak	No	Yes	
2	11.1F-77R	Yes	Weak	No	Yes	
2	77R-13F	Yes	None	No	No	Centromere
2	13F-5.1R	Yes	Weak	Yes	Yes	
2	5.1R-46.1F	No	Strong	Yes	No	
2	46.1F-23.2F	Yes	Strong	Yes	No	
2	23.2F-40F	Yes	Strong	Yes	No	Telomere
3	30R-18R	Yes	Strong	Yes	No	Telomere
3	18R-34.2R	Yes	Strong	Yes	No	
3	34.2R-2.2F	Yes	Weak	Yes	Yes	
3	2.2F-0R	No	Weak	Yes	No	Centromere
3	0R-7.1R	Yes	Strong	Yes	Yes	

3	7.1R-34.1F	Yes	Strong	Yes	No	
3	34.1F-2.1R	Yes	Strong	Yes	Yes	
4	59R-43F	No	Strong	Yes	No	Telomere
4	43F-60F	No	Strong	Yes	No	
4	60F-72R	No	Strong	Yes	No	
4	72R-4R	Yes	Strong	Yes	No	
4	4R-5.2R	No	None	No	No	Centromere
4	5.2R-99R	Yes	Weak	No	Yes	
4	99R-42F	No	Weak	No	No	
4	42F-16R	Yes	Strong	Yes	No	Telomere
5	10F-27F	No	Strong	Yes	No	Telomere
5	27F-69R	No	Weak	Yes	No	
5	69R-24.2R	Yes	Weak	Yes	Yes	
5	24.2R-48R	Yes	None	No	Yes	
5	48R-25F	No	None	No	No	Centromere
5	25F-49.2F	Yes	Strong	Yes	Yes	
5	49.2F-26.2F	Yes	Strong	Yes	Yes	
5	26.2F-90F	No	Strong	Yes	No	Telomere
6	74F-62F	No	Strong	Yes	No	
6	62F-50R	No	Strong	Yes	No	
6	50R-118R	No	Strong	Yes	No	
6	118R-51F	No	Strong	Yes	No	
6	51F-15.1R	No	Strong	Yes	No	
6	15.1R-46.2R	No	Weak	Yes	No	
6	46.2R-23.1R	Yes	Weak	No	Yes	
6	23.1R-22R	No	Weak	No	No	
6	22R-36R	Yes	None	No	No	Centromere
6	36R-32.2F	Yes	Weak	No	Yes	
6	32.2F-14R	No	Weak	No	No	
6	14R-85R	No	Strong	Yes	No	Telomere
7	123R-7.3R	Yes	Strong	Yes	No	Telomere
7	7.3R-1.3R	Yes	Strong	Yes	Yes	
7	1.3R-144F	Yes	Strong	No	Yes	
7	144F-56F	No	Strong	No	No	
7	56F-87F	No	Strong	Yes	No	
7	87F-125F	Yes	Strong	Yes	No	
7	125F-106F	Yes	Strong	No	No	
7	106F-15.3R	No	Strong	No	No	
7	15.3R-80R	Yes	Strong	No	Yes	
7	80R-9R	No	Weak	No	No	
7	9R-8.1R	No	None	No	No	Centromere
7	8.1R-8.2R	Yes	Weak	Yes	Yes	
7	8.2R-28.1R	Yes	Weak	No	Yes	
7	28.1R-8.3F	Yes	Weak	Yes	Yes	
7	8.3F-15.2F	No	Weak	Yes	No	
7	15.2F-92F	Yes	Strong	Yes	Yes	

7	92F-45F	Yes	Strong	Yes	Yes	
7	45F-103F	No	Strong	Yes	No	Telomere
8	83R-33R	Yes	Strong	Yes	Yes	
8	33R-21.1R	Yes	Weak	Yes	Yes	
8	21.1R-104R	No	Weak	Yes	No	
8	104R-3.1R	Yes	Weak	Yes	Yes	
8	3.1R-41F	Yes	Weak	Yes	Yes	
8	41F-19R	No	Weak	Yes	No	Centromere
8	19R-63R	No	Strong	Yes	No	
8	63R-79R	No	Strong	Yes	No	
8	79R-76F	No	Strong	Yes	No	Telomere
9	3.2R-84R	Yes	Strong	Yes	No	Telomere
9	84R-1.2F	Yes	Strong	Yes	Yes	
9	1.2F-44R	Yes	Weak	No	Yes	
9	44R-65.1F	Yes	Weak	No	Yes	
9	65.1F-31.2R	No	Weak	No	No	Centromere
9	31.2R-12.2F	Yes	Weak	No	Yes	
9	12.2F-31.1F	Yes	Weak	No	Yes	
9	31.1F-12.1R	Yes	Weak	Yes	Yes	
9	12.1R-71R	No	Strong	Yes	No	
9	71R-88R	Yes	Strong	Yes	No	
9	88R-81R	Yes	Strong	Yes	No	
10	64F-119F	No	Strong	Yes	No	Telomere
10	119F-39F	No	Strong	Yes	No	
10	39F-68F	Yes	Strong	Yes	No	
10	68F-6R	No	Strong	Yes	No	
10	6R-86R	No	Weak	Yes	No	Centromere
10	86R-17R	Yes	Weak	Yes	Yes	
10	17R-75R	No	Weak	Yes	No	
10	75R-70R	Yes	Strong	No	No	
10	70R-65.2F	Yes	Strong	No	Yes	
10	65.2F-126F	Yes	Strong	Yes	Yes	
10	126F-20F	Yes	Strong	Yes	Yes	Telomere

Table S7: Genome size evolution and distribution of recombination in sorghum, rice, and maize.

	Rice	Sorghum	Maize
Genome size (mbp)	420	740	2160
Repetitive DNA (mbp, % total)	168 (40%)	460 (62%)	1770 (82%)
Retroelements (mbp, %)	109 (26%)	400 (54%)	1706 (79%)
Recombination-poor DNA (heterochromatin?) (mbp, %)	63 (15%)	460 (62%)	773 (36%)
Recombination in recombination-poor DNA , cM (% of total)	30 cM (2%)	34 cM (3%)	361 cM* (4.8%)
Gene models in recombination-poor DNA	1717	8477	N. A.
Recombinogenic DNA (euchromatin?) (mbp, %)	309 (73.6%)	252 (34.1%)	1380 (64.1%)
Recombination in recombinogenic DNA, cM (% of total)	1497 cM (98%)	1025 cM (97%)	7047 cM* (95.2%)

S2.7 Organellar sequences.

The *Sorghum bicolor* mitochondria and chloroplast have been previously sequenced and are in Genbank as accessions NC_008360 and NC_008602. Because of the very clean nuclear DNA preparation used here, we did not have enough organelle "contamination" in the shotgun data to recreate both organelles from the WGS set. We did, however, verify that the sequences are identical to those available in Genbank, except for 1 bp in the mitochondrial genome, which was brought to the attention of the owner of the Genbank record.

Insertions of organellar DNA into the nuclear genome of Sorghum bicolor

For the analysis the assembled nuclear genome of Sorghum was compared against the Sorghum plastid genome (EF115542) and the sorghum mitochondrial genome (DQ984518) respectively.

BLASTN was carried out locally using standard settings. We identified all hits longer than 50 bp for further analysis.

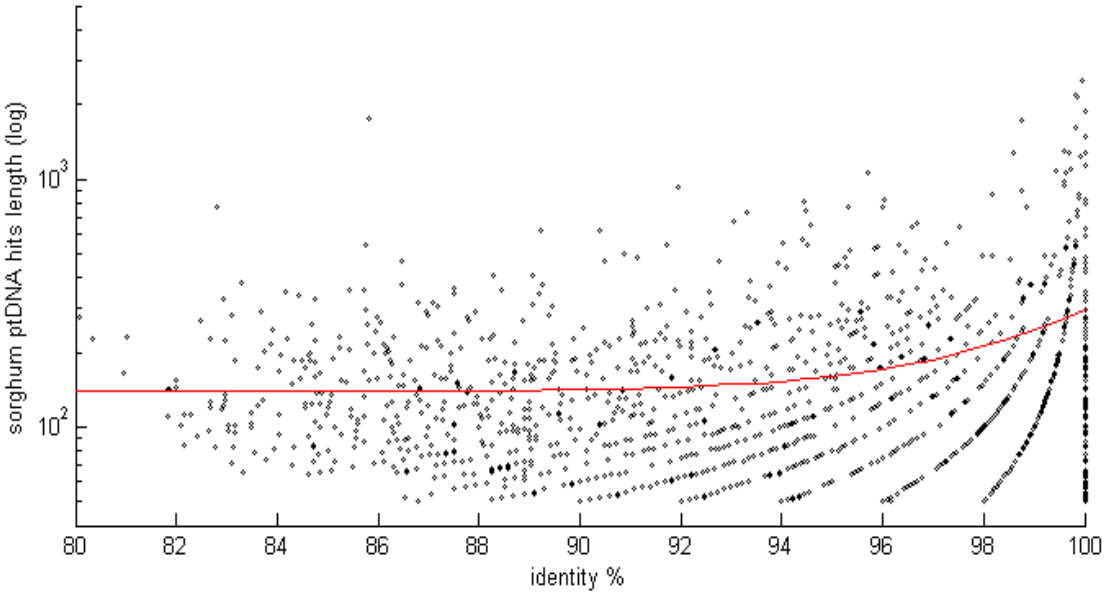
Sorghum plastid DNA vs. Sorghum nuclear genome

1402 insertions derived from the plastid genome have been identified (Table S7).

Table S8: Length and distribution of chloroplast DNA insertions on the Sorghum chromosomes.

Chromosome	***Amt. ptDNA (bp)***	***Number of ptDNA insertions***
1	33084	217
2	30533	183
3	69304	241
4	25431	154
5	15055	77
6	14705	106
7	13579	88
8	19778	117
9	13413	103
10	17132	116

A total of 1,337 insertions detected are shorter than 500bp, with 47 between 0.5 and 1 kb, 15 between 1 and 2 kb, and only 3 exceeding 2 kb with the largest being 2483 bp. As illustrated below, sequence identity between the organellar DNA and the nuclear insertion is greater for longer inserts.



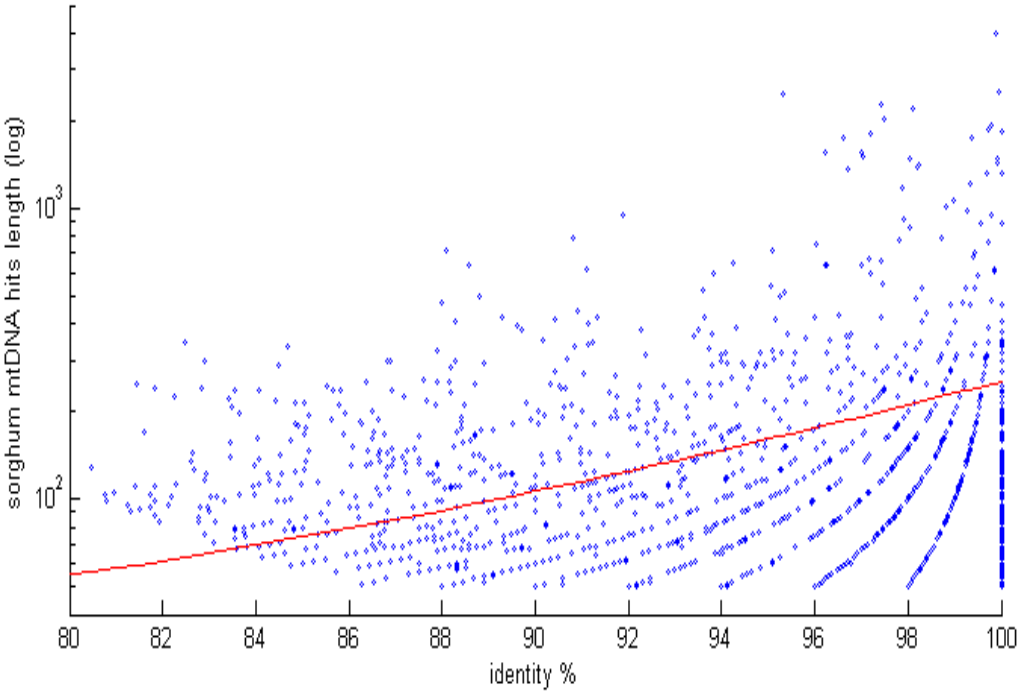
Sorghum mtDNA vs. Sorghum nuclear DNA

A total of 2125 insertions derived from sorghum mtDNA have been detected. Similar to the findings for the chloroplast insertions, 2,052 are less than 500 nt in length, with 46 between 0.5 and 1 kb, 21 between 1 and 2 kb, 5 between 2 and 3 kb, and 1 exceeding 3 kb (3973 bp).

Table S9: Length and distribution of mitochondrial DNA insertions on the Sorghum chromosomes.

Chromosome	*Amt. mtDNA (bp)*	*No. mtDNA insertions*
1	37777	350
2	43423	239
3	59613	281
4	31144	206
5	19005	157
6	20341	200
7	19918	159
8	21315	151
9	27313	165
10	23852	217

Similar to observations for insertions derived from chloroplasts mitochondrial insertions show a pronounced correlation between insertions length and sequence conservation which illustrates an existing elimination mechanism.



S2.8 CpG Island Detection.

The EMBOSS (http://emboss.sourceforge.net/) program, *newcpgreport*, was used to call CpG islands with these parameters: CG observed/expected ratio ($CG_{o/e}$) > 1.2; %[C+G] > 50.00; Length > 200; window size = 100. The actual $CG_{o/e}$ in the sorghum genome is 0.691 assuming an expected CG frequency of 0.125%. EMBOSS output was parsed with in-house Perl scripts.

S3. Repeat identification and characterization

Known repeats were identified with RepeatMasker (www.repeatmasker.org) with a database of grass repeats (mips-REdat_6.2_Poaceae.lib) that contains previously known sorghum-specific LTR retrotransposons and those newly identified from the genome assembly as described below in S3.1. A summary of the repetitive DNA content can be found in Table S10.

S3.1 Identification of LTR-retrotransposons

De novo searches for LTR retrotransposons were performed with LTR_STRUCT (pmid 12584121) on the 10 sorghum chromosomes and all unassembled contigs > 10 kb. The program yielded 10,126 full-length LTR retrotransposon candidate sequences, which were checked for the typical retrotransposon protein domains (GAG, PR, INT, RT) by a

HMMer (http://hmmer.janelia.org) search against respective pfam hmm models. 8071 (80%) of the candidate sequences remained after a quality check and overlap removal. The main quality criteria are the existence of at least one typical retrotransposon protein domain and a simple sequence and tandem repeat content<=35%. According to their protein signatures 2985 (37 %) could be assigned to the gypsy (PR-RT-INT) and 724 (9%) to the copia (PR-INT-RT) LTR superfamily, the remaining 4362 (54%) are temporarily unclassified until the evaluation of further cluster analyses.
A nonredundant set of 7643 quality checked LTR retrotransposons was added to mipsREdat (mips.gsf.de/proj/plant/webapp/recat/), a plant repeat element database, used for the homology based repeat masking and annotation (S3.6).

The insertion age of full length LTR-retrotransposons was determined from the evolutionary distance between 5' and 3' soloLTR derived from a ClustalW alignment of the two solo LTRs by the Kimura two-parameter method (emboss distmat, http://emboss.sourceforge.net/). For the conversion of distance to insertion age, a substitution rate of 1.3E-8 mutations per site per year was used (pmid 15240870). An additional 4,192 full length LTRs were detected by the similarity search (S3.6), thus adding up to 11,352 full length sorghum LTRs on the assembled sorghum chromosomes, for which the insertion age could be calculated (Fig. S2).

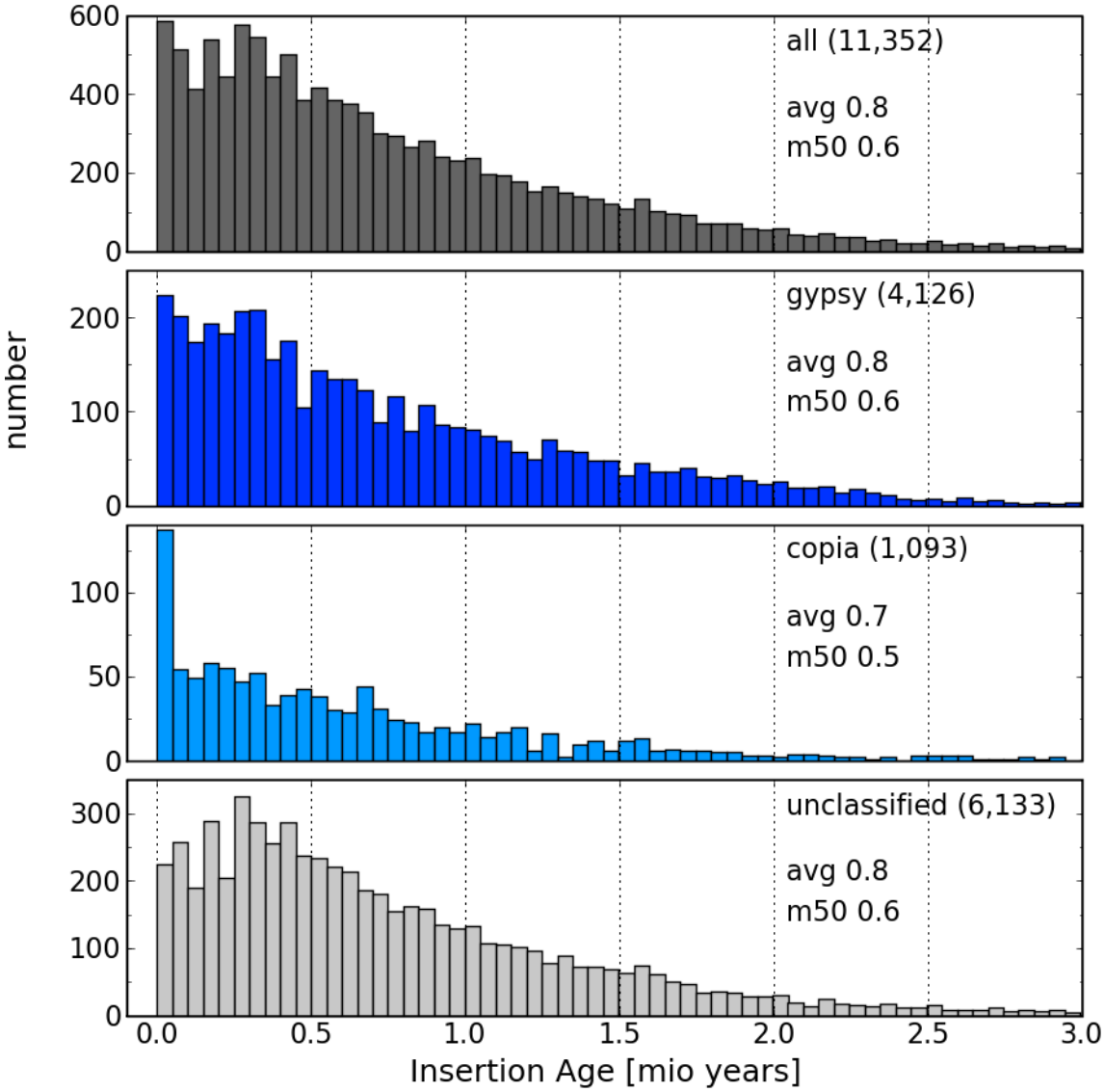


Figure S3: Timing of LTR-Retrotransposon Insertions. The insertion age of full length LTR-retrotransposons was determined from the divergence of left and right solo LTR as described (S 3.1). The bars represent bins of 0.05 million years (mya). The more or less constant increase starting about 5 mya is the outcome of two opposite forces: 1. element insertion and 2. removal and deterioration**.** Copia elements, which amount only to ¼ of the copy numbers of gypsy-type elements show a very recent exceptional number increase.

S3.2 MITEs

Full-length Miniature Inverted-repeat Transposable Elements (MITEs) were identified based on their inverted repeat structure and their 2 and 3 bp target site duplications for *Stowaway* and *Tourist* MITEs, respectively. The initial set identified in this way was used for multiple sequence alignments in order to identify families and to construct consensus sequences for all of them. The consensus sequences were used for a BLAST survey to identify all elements, including fragments.

S3.3 Masking based on over-represented 16-mers

We also performed an *ab initio* search for repetitive elements as follows. The shotgun reads were scanned for 16-mers that occurred in the dataset more than 100 times. At 8.5X nominal coverage, such sequences are ~12-fold overrepresented relative to a naïve expectation from a non-repetitive genome, and include both simple sequence repeats (microsatellites) as well as other highly represented sequences (e.g., 16-bp fragments of minisatellites, retroelements, etc.) Note that many particular instances of a repetitive element in the genome have short stretches that are unique in the genome (e.g., due to mutations after retroelement insertion); in part, this variation between repeats allows assemblies of repetitive regions to be made. To represent repetitive regions, we first grouped blocks of overlapping over-represented 16-mers if they spanned more than 100 bp, and then grouped these blocks if the gaps between them were shorter than 140 bp, a spacing that was empirically determined.

There is good agreement between the over-represented 16-mer masking and masking based on known repeats, and for gene predictions we use the masking from known repeats as being more precise and less prone to masking recently expanded protein-coding gene families.

Table S10: Repeat composition and major components of the sorghum genome in comparison to rice and maize

	Os	**Sb**	**Zm**
	% of genome bp		
Class I: Retroelement	**25.78**	**54.52**	**79.44**
LTR Retrotransposon	**23.47**	**54.43***	**75.08***
Ty1/copia	2.47	5.18	21.75
Ty3/gypsy	12.03	19.00	37.73
unclassified LTR	8.98	30.25	15.59
non-LTR Retrotransposon	**1.24**	**0.04**	**0.35**
LINE	0.80	0.04	0.34
SINE	0.45	0,00	0.01
unclassified retrotroelement	1.07	0.05	4.02
Class II: DNA Transposon	**13.67**	**7.46**	**2.68**
DNA Transposon Superfamily	**7.04**	**4.79**	**0.92**
CACTA superfamily	3.43	4.69*	0.47
hAT superfamily	0.52	0.02	0.10
Mutator superfamily	1.81	0.06	0.15
Tc1/Mariner superfamily	0.02	0.00	0.00
PIF/Harbinger	0.00	0.02	0.08
unclassified	1.26	0.00	0.12
MITE	**5.24**	**1.74***	**0.32**
Stowaway	1.74	0.19	0.03
Tourist	1.50	0.94	0.08
unclassified MITE	2.00	0.61	0.21
Helitron	**0.33**	**0.81***	**1.31***
unclassified DNA transposon	1.06	0.12	0,12
Transposon DNA	**39.5**	**62.0**	**82.1**
Coding space	**33.0**	**14.3**	**7.5**
Unassigned space incl. regulatory seq	**27.6**	**23.7**	**10.4**

The transposon space of sorghum (Sb), maize (Zm; 100 random BACs {pmid 16339807}) and rice (Os; TIGR 5 assembly) was annotated as described in section 3.6. Asterisks (*) mark element types for which an additional *de novo* detection was carried out to complement the homology based approach.

Table S11: Repeat composition by type

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Table S12: Lineage specificity of transposons

Clade	Number	Percent of number	Nucleotides	Percent of repetitive nucleotides	Percent of Genome
Panicoideae	290,052	99.99	383,797,790	99.98	51.97
Andropogoneae	290,042	99.99	383,795,558	99.98	51.97
Erianthus	61	0.02	9,526	0.00	0.00
Sorghum	281,416	97.01	380,371,731	99.09	51.50
Zea	8.565	2.95	3,414,301	0.89	0.46
Paniceae	19	9	2,232	0.00	0.00
Setaria	10	0	2,232	0.00	0.00

Table S13. Repetitive content per chromosome

Chromosome	Total length (including spanned gaps and centromere) [Mb]	Percent G+C	Percent A+T	Repeat-masked sequence (not including unassembled regions)[%]	Percent masked sequence based on regions dominated by over-represented 16-mers	Gaps in sequence assembly (including centromere) [%]
1	73.8	44.9	55.1	44.6	40.2	0.8
2	77.9	44.6	55.4	61.6	54.7	1.3
3	74.4	44.9	55.1	59.3	53.4	0.9
4	68.0	44.7	55.3	57.3	51.1	0.4
5	62.4	43.5	56.5	66.0	57.9	0.8
6	62.2	44.8	55.2	67.4	61.1	0.6
7	64.3	44.1	55.9	67.4	60.1	0.7
8	55.5	43.5	56.5	66.1	58.5	1.7
9	59.6	44.4	55.6	63.0	56.4	1.1
10	61.0	44.5	55.5	61.6	55.2	0.5
Total mapped sequence	**659.2**	**44.4**	**55.6**	**61.4**	**54.5**	
Unmapped sequence					**81.2**	

S3.4 CACTA Search Strategy

A program was developed (T. Wicker) that specifically searches for the typical CACTA TIR pattern. Most CACTA elements have arrays of direct and inverted repeat units (about 20-40 bp per unit) in their terminal regions. They can be nicely visualised with a DotPlot of a CACTA sequence against itself (Figure S3).

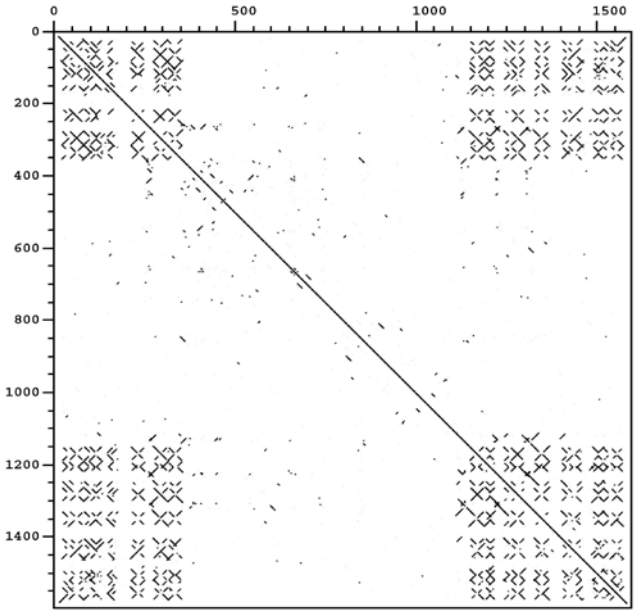


Figure S4. Dotplot of a small non-autonomous CACTA element sequence compared with itself. The terminal repeat arrays cause a characteristic pattern. This non-autonomous element is a deletion derivative that has lost all of its coding regions.

Search step 1

The program specifically searches for strings (5-7 bp) which occur in forward and reverse orientation within a region of ~500 bp. If that is found, it searches for a second region 1 – 12 kb away from it. If that is found, it looks if the element is flanked by the characteristic palindromic CACTA/G...C/TAGTG termini flanked by a 3 bp TSD.

Search Step 2

Since this pattern occurs more than 8,000 times in the sorghum genome, a second more stringent step was added to exclude chance occurrences. Using the Smith-Waterman algorithm, the 2 termini of each candidate were aligned and checked for the presence of imperfect terminal inverted repeats of at least 10 bp in size. This resulted in several hundred complete elements. From these, all known elements were identified by BLAST against the known CACTA elements.

The novel elements were BLASTed against one another to identify families. False positives that were still the product of chance occurrence were then selected out by hand. The result is 95 new CACTA families. Most of them have a moderate copy number. To date, only complete elements have been described.

S3.5 Helitron annotation

We calibrated the helitron finder of Du et al [23] based on helitrons predicted in a 1MB alignment between *S. bicolor* and *S. propinquum* and applied it to the whole genome. Estimates of helitron content in maize were made using the appropriate maize calibrations. [23]

Table S14: Helitrons in sorghum and maize

	Sorghum			**Maize**	
	all	chromosomes	unmapped	100 BACs	2 contiguous sequences
Genomic sequence [Mb]	738.5	659.2	79.3	14.4	14.4
# helitrons	1355	1017	338	22	29
sum of helitrons [Mb]	**7.2**	5.0	2.1	0.19	0.25
average length [kb]	**5.3**	5.0	6.3	8.6	8.8
median length [kb]	**3.6**	3.4	4.0	6.9	6.6
% in genome	**0.97**	0.77	2.68	**1.31**	**1.76**

S3.6 Tandem repeats

Tandem Repeats where detected by the program Tandem repeats finder (tandem.bu.edu/trf/trf.html) {pmid 9862982} with default parameters (2,7,7,80,10,50,500). Depending on monomer length the tandem repeats where classified as microsatellites (2-6 bp), minisatellites (7-100 bp) or satellites (> 100 bp). Overlaps where removed by collapsing all trf annotations on the genomic sequence. The tandem repeat content is summarized in Table S15. A list of SSR and VNTR loci potentially useful as DNA markers is provided in Supplementary List 1.

Table S15: Tandem repeats

	#	# %	% of tandem repeat bp	perc of genome
Tandem Repeats	109,039	100	100.0	3.13
Microsatellite	15,194	14	4.4	0.14
Minisatellite	80,932	74	31.8	1.00
Satellite	12,913	12	63.8	2.00
Cen 38	4,229	4	46.9	1.47

S3.7 Repeat annotation and data integration

Diverged transposons and their fragments were detected with RepeatMasker Open-3-1-8 {www.repeatmasker.org} using a customized grass repeat library (mips-REdat_6.2_Poaceae, 15665 sequences, 98.9 Mb) which contained the newly identified sorghum LTR-retrotransposons (S3.1) and MITEs (S3.2) in addition to a non redundant set of known grass transposons from the following sources: TREP (wheat.pw.usda.gov/ITMI/Repeats/), RetrOryza (www.retroryza.org) {pmid 17071960}, TIGR plant repeats databases (www.tigr.org/tdb/e2k1/plant.repeats/){pmid 14681434} and RepBase (www.girinst.org) {pmid 16093699}.

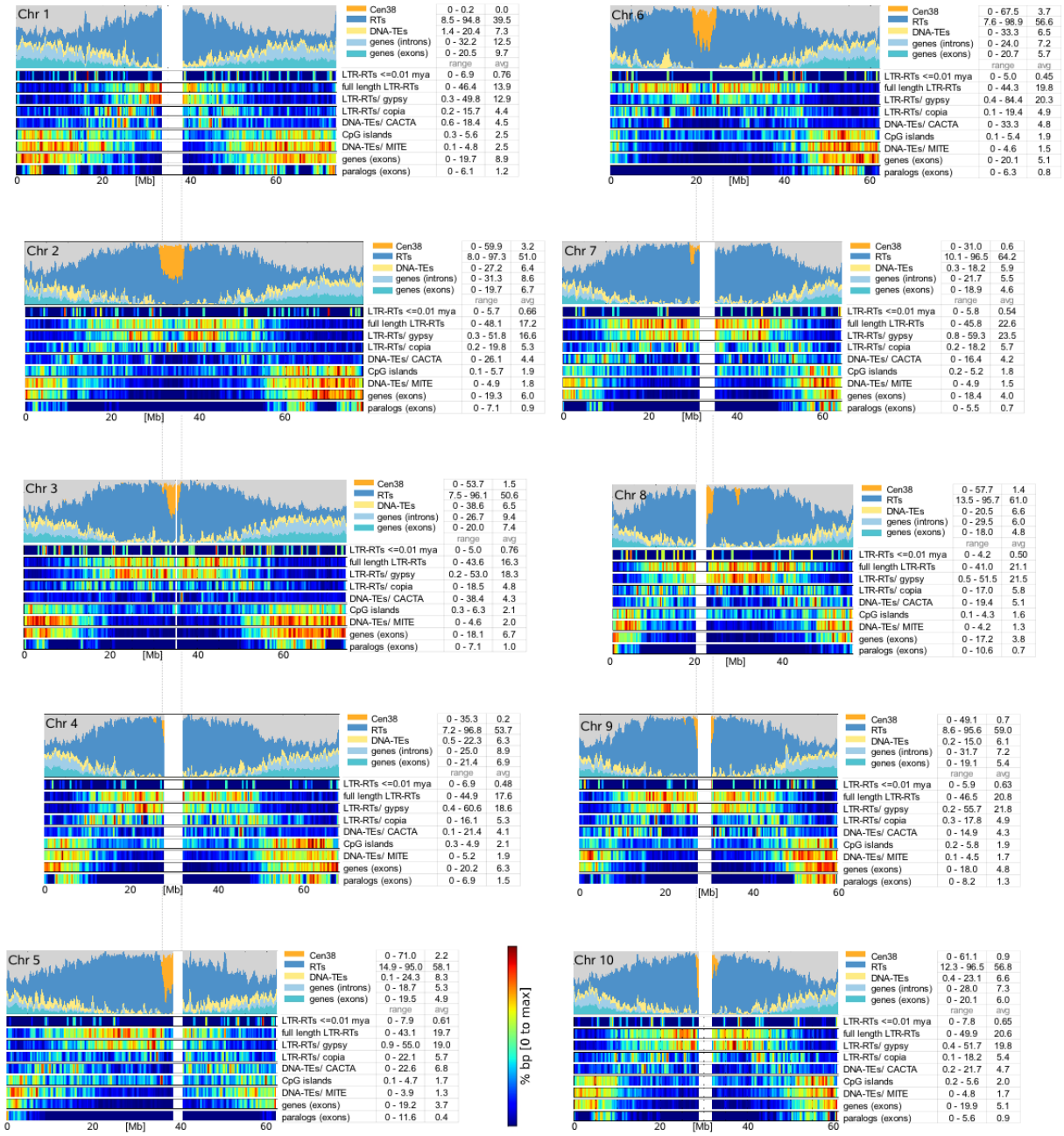
The integration of the specialized transposon data for LTRs (S3.1), MITEs (S3.2), CACTAs (S3.3) and Helitrons (S3.4) into a final consolidated repeat annotation was carried out with modules from the MIPS ANGELA pipeline (**A**utomated **N**ested **G**enetic **El**ement **A**nnotation). Overlapping repeat annotations frequently occur in repeat rich genomes. They are caused by highly similar regions shared by different transposons or by composite elements in the repeat libraries, e.g. LTR retrotransposons with CACTA inserts. The annotation overlaps were handled in a priority based approach. High confidence expert annotations are assigned first, and overlapping elements with lower priority are either truncated, fragmented or skipped, depending on adjustable parameters for overlap percent and minimum rest length. The assignment order within one priority group is defined by descending homology score or element length. For sorghum all elements overlapping > 80% of their length to higher priority elements were discarded, the minimum rest length after truncation was 30 bp, and the following priority order was used: 1. CACTA DNA transposons; 2. MITEs; 3. full length LTR

retrotransposons; 4. RepeatMasker annotation. A summary of the repetitive content can be found in Table S10, together with a comparison to rice and maize.

The distribution of elements along the chromosomes was calculated from a sliding window of 0.5 Mb with 0.1 Mb overlaps (**Figure S5**). For each window the percent bp coverage of the respective element type was calculated with the number of non-N basepairs as denominator. Windows with > 60% N bp were not used and are depicted in white.

Figure S5: Genomic landscape of sorghum.
The stacked barcharts along each chromosome show the proportional distributions of the main DNA element types: retrotransposons, genes split into exons and introns, DNA transposons and the cen38 centromeric tandem repeat. The gray color represents the so far unassigned space and includes regulatory regions. The observation that introns often contain transposons is not displayed in the barcharts, because of its absolute small value of ~1.5% of the genome content. The y-axis goes from 0 to 100 %bp, and the x-axis consists of 0.5 Mb sliding windows with a 0.1 Mb shift. Heatmap tracks visualize the distribution of specific elements (or families) and their correlations. The scale is different for each track, ranging from 0 (blue color) and to the maximum observed number (red color) given in the accompanying tables under the range field. The overall distribution pattern is similar for all chromosomes: cen38 makes up ~ 50% of the centromeric regions, which are thought to have been largely sequenced for chrs. 2, 3 and 6. The high retrotransposon content in the pericentromeric regions gradually decreases towards the gene rich chromosome ends. Gene and retrotransposon densities are negatively correlated, but DNA transposons (especially MITES) co-occur with genes. The short arm of chr. 6 is an exception, largely missing the gene rich region, with no paralogs and a relatively high retrotransposon content, giving the impression of a truncation. Such a truncation would necessarily have been ancient, as the corresponding rice chromosome shows similar gene and repeat distribution.

A high-resolution version of the figures below is also included as Supplementary Image 1, that permits zooming in on a specific region of interest.



S4. Protein coding gene annotation

S4.1 Structural gene calls in the Sorghum genome

Protein-coding genes were derived from the consensus of several sources of evidences as well as *ab initio* predictions. First, TIGR rice transcript assemblies [24] were mapped to

the repeat-masked Sorghum genome sequences applying GenomeThreader [25] and a maize splice site model. Optimal spliced alignments (OSAs) of assemblies and ESTs of the following monocot species have been included: *Allium cepa*, *Ananas comosus*, *Avena sativa*, *Brachypodium distachyon*, *Curcuma longa*, *Hordeum vulgare*, *Oryza sativa*, *Saccharum officinarum*, *Secale cereale*, *Sorghum bicolor*, *Sorghum halepense*, *Sorghum propinquum*, *Triticum aestivum*, *Zea mays*, and *Zingiber officinale*.
We also generated OSAs as well as BlastX alignments for a reference set of proteins consisting of the SWISSPROT database [26] and proteomes of *Arabidopsis thaliana* (TAIR6 version; [27]), *Saccharomyces cerevisiae* [28] and rice [29]. For each OSA, possible reading frames of size ≥50 amino acids were collected as candidates for gene models. In addition, we identified gene models on repeat masked genomic sequences by *ab initio* methods (Fgenesh++, GeneID, GenomeScan).

Next, we applied Jigsaw [30] as statistical combiner of all supporting information from the first analysis round described above. A decision tree has been trained on a set of 987 gene models that were edited by human supervision in the Apollo Genome Browser [31]. All models, including those obtained from the first analysis series, were scored by Blastp against the UniREF90 protein database and for each locus the best fitting model, i.e. the model with the highest bit score, has been selected.

The models were used as input for the PASA pipeline [32] in order to (i) predict UTRs using maize, sorghum and sugarcane ESTs, (ii) identify possible alternative splicing patterns, and (iii) to fit all predicted models to the splice sites suggested by EST evidences of closely related species. Besides complete gene models, we also included candidate (partial) genes that lack a start and/or stop codon. Note that partial gene models may result from several, not mutually exclusive reasons: (i) sequencing or assembly errors may hinder both *ab initio* and homology based predictors to deduce a correct ORF; (ii) transposon activity may have lead to truncated genes or pseudogenes; (iii) insufficient evidence from *ab initio* predictions or EST matches may build and support only incomplete gene models.

S4.2 Gene identifiers

We adopted a gene nomenclature convention based on the time-tested approaches used by the Arabidopsis and rice communities (Eva Huala, TAIR, private communication). Each protein-coding gene locus is assigned a unique identifier of the form "SbXX%YYYYY" where

- "Sb" indicates *Sorghum bicolor*.
- "XX" is a two digit numerical chromosome identifier (01-10) or four digit scaffold identifier (0010-3326)
- The delimiter "%" is either "g" for chromosomally mapped sequences or "s" for scaffolds
- YYYYY is a unique five digit numerical code.

In the initial assignment of locus identifiers, genes are assigned numbers starting from YYYYY=00200 at the start of each assembled sequence, and incrementing by 10. Spans

longer than 100 kb between initially annotated loci are represented by a skip of 200. Thus in the initial assignment, the numerical code corresponds to chromosomal position.

As additional data is generated, the initial chromosomal assembly of *Sorghum bicolor* described here is likely to accumulate small improvements that may result in (1) local rearrangements of modest numbers of genes, including gene model fusions and fissions, (2), placement onto chromosomes of currently unmapped protein-coding genes found in the scaffolds (most if not all of which are centromeric), (3) corrections will be made to predicted gene structures, and (4) new genes will be discovered.

In future releases, the following conventions will be used:

- Revisions and alternate splice isoforms of the protein-coding transcripts at a given locus SbXX%YYYYY will be assigned decimals as in SbXX%YYYYY.1, .2., .3., etc.
- Locus identifiers will be preserved if gene structure corrections are made, as long as the mapping from old to new is unambiguous.
- If a locus is deemed to have inadvertently joined two (or more) adjacent loci, the original locus identifier will be retired and the two nearby new numbers assigned. Note that in some cases this may result in non-monotonic increase of the identifiers along the chromosome.
- as the remaining scaffolds are mapped to chromosomes, loci on these scaffolds will be reassigned new Sb gene identifiers reflecting their appropriate chromosome and position, and the original SbXXsYYYYY number will be retired from use but noted as synonyms. Fewer than 700 predicted genes fall on these unmapped scaffolds.
- In subsequent assemblies, Sbi identifiers will be preserved for genes that unambiguously map forward.

S4.3. Tandem gene clusters in sorghum

Tandem expansions were defined as all sets of peptides with a pairwise Blastp alignment of e-value better than 1e-25 and two or less intervening genes. Characteristics of the largest tandem gene clusters of 8 or more genes in sorghum are briefly summarized below.

First gene	#	4DTV min	4DTV max	majority PFAMs	pfam def
Sb03g028560.1	15	0.042	0.568	PF00067	Cytochrome P450
Sb05g027740.1	14	0.091	0.500	PF03514	GRAS family transcription factor
Sb05g019890.1	14	0.000	0.545	PF02797	Chalcone and stilbene synthases
Sb02g031700.1	14	0.000	0.538	PF02519	Auxin responsive protein
Sb07g024600.1	13	0.000	0.412	PF03087	Arabidopsis protein of unknown function
Sb07g026660.1	13	0.067	0.556	PF00651	BTB/POZ domain
Sb04g003800.1	12	0.101	0.417	PF00560	Leucine Rich Repeat
Sb01g030930.1	12	0.083	0.590	PF00043	Glutathione S-transferase
Sb06g029690.1	11	0.125	0.609	PF07714	Protein tyrosine kinase

Sb02g035420.1	11	0.145	0.542	PF07714	Protein tyrosine kinase
Sb07g001850.1	11	0.000	0.333	PF03087	Arabidopsis protein of unknown function
Sb06g029520.1	11	0.053	0.350	PF01370	NAD dependent epimerase/dehydratase
Sb02g040660.1	11	0.063	0.538	PF00141	Peroxidase
Sb01g030780.1	11	0.163	0.563	PF00043	Glutathione S-transferase
Sb01g029230.1	10	0.030	0.338	PF03330	Rare lipoprotein A (RlpA)-like
Sb05g006750.1	10	0.000	0.034	PF00023	Ankyrin repeat
Sb01g039430.1	10	0.044	0.309	PF00012	Hsp70 protein
Sb05g027220.1	9	0.000	0.444	PF00560	Leucine Rich Repeat
Sb08g022370.1	9	0.013	0.475	PF00314	Thaumatin family
Sb06g022410.1	9	0.037	0.600	PF00232	Glycosyl hydrolase family 1
Sb03g045780.1	9	0.050	0.511	PF00043	Glutathione S-transferase
Sb05g005550.1	9	0.000	0.102	.	
Sb03g027430.1	8	0.000	0.384	PF08370	Plant PDR ABC transporter associated
Sb10g029930.1	8	0.000	0.297	PF07893	Protein of unknown function (DUF1668)
Sb05g004680.1	8	0.000	0.448	PF07762	Protein of unknown function (DUF1618)
Sb03g001810.1	8	0.063	0.508	PF00657	GDSL-like Lipase/Acylhydrolase
Sb10g026720.1	8	0.051	0.500	PF00651	BTB/POZ domain
Sb07g005230.1	8	0.000	0.477	PF00190	Cupin
Sb02g025250.1	8	0.176	0.573	PF00060	Ligand-gated ion channel
Sb10g030620.1	8	0.071	0.333	.	
Sb02g001210.1	8	0.064	0.422	.	
Sb02g034100.1	8	0.043	0.440	.	
Sb05g005756.1	8	0.059	0.231	.	
Sb03g041190.1	8	0.054	0.727		

S4.4 Sorghum miRNA gene annotation

We annotated sorghum microRNAs in two steps. First, we mapped the existing sorghum miRNA entries of miRBase release 11 [33] to the sorghum genome. Second, we used rice miRNAs from miRBase release 11 to annotate new sorghum miRNA genes since very recently several deep sequencing projects reported many new rice miRNAs. After a rice miRNA was mapped to the sorghum genome, the surrounding sequence was checked for hairpin structures. Those loci which fulfilled miRNA precursor secondary structures were annotated as sorghum miRNA genes. We have annotated 149 miRNA genes in the sorghum genome.

Natural antisense miRNAs (nat-miRNAs) were recently identified in monocots. They are located at the antisense strand of their target genes and contain long introns in their precursor sequences. Three sbi-miR444 precursors were mapped to the Sorghum genome. Interestingly, one sbi-miR444 locus produces two precursors due to exon skipping. The targets of miR444 are MADS box proteins, important regulators of plant development.

The miR821 family has five members in Sorghum. Their precursor sequences are highly similar (~80% nucleotide similarity) to the rice ortholog miRNA precursors but the

mature miRNA sequences are not identical. There are one or two nucleotide differences between the osa-miR821 sequence and the sbi-miR821 sequences.

Table S16: miRNAs present in the sorghum genome

miRNA gene family	Known miRNA genes*	Paralogous miRNA genes	Total miRNA genes	miRNA genes found in cluster** (# of clusters)
miR156	5	4	9	2 (1)
miR159	2	0	2	
miR160	5	1	6	
miR162	0	1	1	
miR164	3	2	5	
miR166	7	4	11	
miR167	7	3	10	
miR168	1	0	1	
miR169	7	7	14	2 (1)
miR171	6	5	11	
miR172	4	1	5	
miR319	1	1	2	
miR390	0	1	1	
miR393	1	1	2	
miR394	1	1	2	
miR395	5	7	12	11 (3)
miR396	3	2	5	
miR397	0	1	1	
miR399	9	1	10	
miR408	0	1	1	
miR437	0	23	23	
miR444	0	3	3	
miR528	0	1	1	
miR529	0	1	1	
miR821	0	5	5	
miR1432	0	1	1	
miR1435	0	2	2	
miR1436	0	1	1	
miR1439	0	1	1	
Total	67	82	149	15 (5)

* Based on miRBase v11
** Using clustering length of 500 nucleotides

Table S17: Position of known sorghum miRNAs in the genome

miRNA	Precursor Length	Chromosome	Precursor Start	Precursor End	Strand
sbi-MIR156a	84	4	5373547	5373630	[-]
sbi-MIR156b	84	3	3473048	3473131	[-]
sbi-MIR156c	95	3	3473369	3473463	[-]
sbi-MIR156d	125	2	62836722	62836846	[-]

sbi-MIR156e	123	10	55009872	55009994	[+]
sbi-MIR159	226	3	8194328	8194553	[-]
sbi-MIR159b	253	3	1225082	1225334	[-]
sbi-MIR160a	84	4	4236169	4236252	[-]
sbi-MIR160b	82	10	56834481	56834562	[+]
sbi-MIR160c	83	7	2730531	2730613	[+]
sbi-MIR160d	100	1	5215950	5216049	[-]
sbi-MIR160e	95	2	2925262	2925356	[-]
sbi-MIR164	126	9	39002547	39002672	[-]
sbi-MIR164b	111	4	64881672	64881782	[-]
sbi-MIR164c	153	1	61593529	61593681	[-]
sbi-MIR166a	108	1	17295171	17295278	[-]
sbi-MIR166b	72	1	7426521	7426592	[+]
sbi-MIR166c	94	1	69265260	69265353	[-]
sbi-MIR166d	87	4	63283312	63283398	[-]
sbi-MIR166e	151	2	61439831	61439981	[+]
sbi-MIR166f	139	4	64225347	64225485	[-]
sbi-MIR166g	134	4	64225078	64225211	[-]
sbi-MIR167a	96	1	4354681	4354776	[+]
sbi-MIR167b	198	1	7272352	7272549	[+]
sbi-MIR167c	133	10	56170660	56170790	[+]
sbi-MIR167d	148	2	4993335	4993482	[-]
sbi-MIR167e	179	8	51954679	51954857	[+]
sbi-MIR167f	179	1	26225027	26225205	[+]
sbi-MIR167g	123	3	64088364	64088486	[-]
sbi-MIR168	106	4	2246316	2246421	[-]
sbi-MIR169a	91	3	10825168	10825258	[+]
sbi-MIR169b	102	10	55869177	55869278	[-]
sbi-MIR169c	126	6	39830386	39830511	[+]
sbi-MIR169d	125	6	39791164	39791266	[+]
sbi-MIR169f	148	2	64603670	64603817	[+]
sbi-MIR169g	152	2	64606503	64606654	[+]
sbi-MIR169i	169	5	17050323	17050491	[+]
sbi-MIR171a	161	1	7845711	7845871	[-]
sbi-MIR171b	132	7	7609099	7609230	[+]
sbi-MIR171c	109	2	17125729	17125837	[-]
sbi-MIR171d	154	1	71039535	71039687	[-]
sbi-MIR171e	124	6	54609030	54609153	[+]
sbi-MIR171f	119	4	62099903	62100021	[-]
sbi-MIR172a	102	9	58962031	58962132	[-]
sbi-MIR172b	170	3	74241513	74241682	[-]
sbi-MIR172c	119	4	67015298	67015416	[-]
sbi-MIR172e	115	2	14181315	14181429	[-]
sbi-MIR319	249	3	1240163	1240411	[+]
sbi-MIR393	139	3	6521844	6521966	[+]
sbi-MIR394a	110	2	66910962	66911071	[+]
sbi-MIR395a	150	6	58760409	58760558	[+]
sbi-MIR395b	105	6	58761003	58761107	[+]
sbi-MIR395d	104	6	58197343	58197445	[-]
sbi-MIR395e	105	6	58197534	58197638	[-]
sbi-MIR395f	122	6	58196833	58196954	[-]
sbi-MIR396a	125	4	66092514	66092638	[-]
sbi-MIR396b	128	10	4424888	4425015	[+]
sbi-MIR396c	162	4	66085287	66085448	[+]

sbi-MIR399a	137	3	61886672	61886801	[+]
sbi-MIR399b	123	4	9842715	9842828	[-]
sbi-MIR399c	130	9	55682225	55682354	[-]
sbi-MIR399d	205	10	1544093	1544297	[+]
sbi-MIR399e	126	9	55683792	55683917	[+]
sbi-MIR399f	120	10	48048104	48048223	[-]
sbi-MIR399g	125	9	55688228	55688352	[+]
sbi-MIR399h	132	10	48050465	48050596	[+]
sbi-MIR399i	121	6	55042923	55043043	[+]

Table S18: Position of newly detected miRNAs (paralog mapping) in the sorghum genome

miRNA family	Paralog Id	Precursor Length	Chromosome	Precursor Start	Precursor End	Strand
156	sbi-MIR156.p1	128	2	59375957	59376084	[+]
156	sbi-MIR156.p2	89	4	55586941	55587029	[-]
156	sbi-MIR156.p3	94	6	50885046	50885139	[-]
156	sbi-MIR156.p4	121	7	55808117	55808237	[+]
160	sbi-MIR160.p1	110	7	63646187	63646296	[-]
162	sbi-MIR162.p1	127	4	55056602	55056728	[+]
164	sbi-MIR164.p1	160	2	76402755	76402914	[-]
164	sbi-MIR164.p2	205	9	45090953	45091157	[+]
166	sbi-MIR166.p1	185	10	59811964	59812025	[-]
166	sbi-MIR166.p2	86	1	27013404	27013489	[+]
166	sbi-MIR166.p3	108	1	72676204	72676311	[-]
166	sbi-MIR166.p4	103	8	39559277	39559379	[+]
167	sbi-MIR167.p1	90	1	69498654	69498743	[-]
167	sbi-MIR167.p2	157	4	4488304	4488460	[-]
167	sbi-MIR167.p3	132	8	51952391	51952522	[+]
169	sbi-MIR169.p1	109	2	58896558	58896666	[+]
169	sbi-MIR169.p2	98	4	49253706	49253803	[+]
169	sbi-MIR169.p3	123	4	58034693	58034815	[+]
169	sbi-MIR169.p4	97	6	56718764	56718860	[-]
169	sbi-MIR169.p5	97	7	61062640	61062736	[+]
169	sbi-MIR169.p6	92	7	61068027	61068118	[-]
169	sbi-MIR169.p7	93	7	61071181	61071273	[+]
171	sbi-MIR171.p1	77	10	54088668	54088744	[+]
171	sbi-MIR171.p2	81	1	15608735	15608815	[-]
171	sbi-MIR171.p3	90	1	52558149	52558238	[-]
171	sbi-MIR171.p4	108	4	5853293	5853400	[-]
171	sbi-MIR171.p5	87	6	57730663	57730749	[-]
172	sbi-MIR172.p1	88	2	22209593	22209680	[+]
319	sbi-MIR319.p1	179	3	58360214	58360392	[-]
390	sbi-MIR390.p1	179	1	2870994	2871172	[+]
393	sbi-MIR393.p1	84	6	61406228	61406311	[-]
394	sbi-MIR394.p1	77	4	62277173	62277249	[-]
395	sbi-MIR395.p1	73	6	58197025	58197097	[-]
395	sbi-MIR395.p2	68	6	58761182	58761249	[+]
395	sbi-MIR395.p3	81	6	58761343	58761423	[+]
395	sbi-MIR395.p4	97	7	4657883	4657979	[+]

395	sbi-MIR395.p5	87	7	4658066	4658152	[+]
395	sbi-MIR395.p6	80	7	4658236	4658315	[+]
395	sbi-MIR395.p7	84	7	4658542	4658625	[+]
396	sbi-MIR396.p1	95	4	67655140	67655234	[-]
396	sbi-MIR396.p2	189	6	60827025	60827213	[+]
397	sbi-MIR397.p1	91	4	4027097	4027187	[-]
399	sbi-MIR399.p1	93	4	9862936	9863028	[-]
408	sbi-MIR408.p1	205	3	15944292	15944496	[+]
437	sbi-MIR437.p1	190	10	3078189	3078378	[-]
437	sbi-MIR437.p2	171	1	19827219	19827389	[+]
437	sbi-MIR437.p3	175	1	20771120	20771294	[+]
437	sbi-MIR437.p4	136	1	23998927	23999062	[-]
437	sbi-MIR437.p5	101	1	54176829	54176929	[+]
437	sbi-MIR437.p6	174	1	56906489	56906662	[-]
437	sbi-MIR437.p7	176	1	59496276	59496451	[-]
437	sbi-MIR437.p8	170	1	60249197	60249366	[+]
437	sbi-MIR437.p9	172	1	73814238	73814409	[+]
437	sbi-MIR437.p10	172	1	9824853	9825024	[-]
437	sbi-MIR437.p11	170	2	45985879	45986048	[+]
437	sbi-MIR437.p12	174	3	49109184	49109357	[-]
437	sbi-MIR437.p13	182	3	6385582	6385763	[-]
437	sbi-MIR437.p14	175	4	66322673	66322847	[-]
437	sbi-MIR437.p15	188	4	8595515	8595702	[-]
437	sbi-MIR437.p16	160	6	10908271	10908430	[-]
437	sbi-MIR437.p17	180	6	35895917	35896096	[-]
437	sbi-MIR437.p18	176	6	49901016	49901191	[-]
437	sbi-MIR437.p19	170	7	5519054	5519223	[+]
437	sbi-MIR437.p20	160	9	47617262	47617421	[+]
437	sbi-MIR437.p21	164	9	53125183	53125346	[+]
437	sbi-MIR437.p22	114	9	53472511	53472624	[+]
437	sbi-MIR437.p23	173	9	55290467	55290639	[+]
444*	sbi-MIR444.p1		4	59018312	59021783	[+]
444*	sbi-MIR444.p2		4	53728538	53723195	[+]
444	sbi-MIR444.p3	583	6	48663114	48663697	[-]
528	sbi-MIR528.p1	84	1	71476711	71476794	[-]
529	sbi-MIR529.p1	116	4	44092392	44092507	[+]
821	sbi-MIR821.p1	285	1	47182770	47183054	[-]
821	sbi-MIR821.p2	273	2	47490926	47491198	[-]
821	sbi-MIR821.p3	285	2	60192106	60192390	[-]
821	sbi-MIR821.p4	264	6	29609545	29609808	[+]
821	sbi-MIR821.p5	269	9	45069956	45070224	[-]
1432	sbi-MIR1432.p1	184	2	71755920	71756103	[-]
1435	sbi-MIR1435.p1	287	2	60539286	60539572	[+]
1435	sbi-MIR1435.p2	267	7	59967771	59968037	[-]
1436	sbi-MIR1436.p1	475	9	42986140	42986614	[+]
1439	sbi-MIR1439.p1	475	9	5350892	5351366	[-]

*miR444 contains a large intron.

S4.5 Rice annotations used for comparisons

Two annotations of *O. sativa* subsp. *japonica* exist: the manually curated and expressed-sequence focused "RAP2" annotation from the Rice Annotation Project [29] and the automated "TIGR5" annotation from TIGR. These annotations share a core of 26,267 common genes, but the TIGR5 set is significantly larger (41,078 genes total) as it includes genes of unknown function that are likely to be ORFs related to repetitive elements. The RAP2 set is more conservative, and is more likely to omit genes without EST evidence, but also retains a modest number of repeat-derived ORFs. We used both rice annotations in comparisons with sorghum, adopting the following filterings of the full datasets.

S4.5.1 Filtering of RAP2 rice annotation data

Rice genes were downloaded from the RAP Rice Annotation Project (31,439 genes), release 2 from http://rapdb.dna.affrc.go.jp/rapdownload/. In the RAP2 annotation a total of 2,049 of the predicted genes were based on ESTs and not assigned to chromosomal locations in rice. These 2,049 genes were compared to three different independent rice whole genome sequences, the BGI indica WGS genome (http://rise.genomics.org.cn/rice/index2.jsp), the Syngenta, japonica WGS genome (http://rise.genomics.org.cn/rice/index2.jsp) and the IRGSP japonica BAC based genome assembly (http://rapdb.dna.affrc.go.jp/rapdownload/). Of the 2,049 unanchored predicted genes 1,029 showed no similarity to any of the three rice genome assemblies when compared with blastn with e-value threshold e<1e-6. Of the remaining 1,020 genes, 897 were mostly repetitive genes that could have been assigned to multiple locations or were "close" but not perfect matches to existing sequences in the IRGSP. The remaining 123 genes did not match the IRGSP assembly but matched one or both of the WGS assemblies and presumably belonged to regions of the rice genome for which a BAC had not been sequenced.

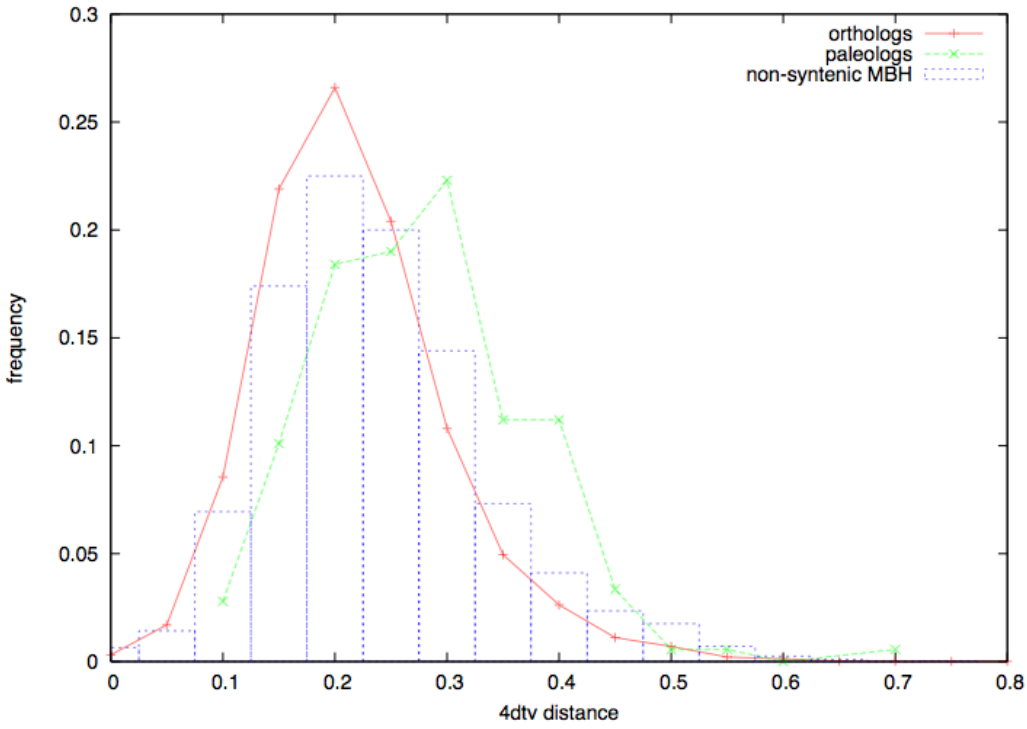
The 1,029 genes that showed no matches to any of the rice genome assemblies were presumed to be largely contaminations of the EST libraries from other sources, likely mostly fungi or yeasts. These 1.029 rice genes were removed from later comparative analysis, leaving 30,410 genes for comparative analysis, with 29,390 of these including positional information.

S4.5.2 TIGR5 gene set for rice

A total of 56,312 longest-at-locus genes were downloaded from TIGR. The 15,424 genes annotated by TIGR as TE-related were removed from the comparison data set, as well as 470 genes that had 10 or more blast (1e-10) hits to other rice genes and contained a transposon-related PFAM domain as the only annotation.

There are 4,006 sbi1_4 genes that have a near mutual-best hit to rice ("C-Score" >0.9) in the set of high-confidence genes without a syntenic ortholog. These orthologs have a

fourfold divergence comparable to typical sorghum-rice orthologous segments, and are evidently bona fide orthologs that have moved from their syntenic position.



S4.6 Protein domains in the Sorghum genome

To gain insight into protein functions and biological processes that are different between *Sorghum* and *Arabidopsis* (TIGR7), rice (RAP2) and poplar, we retrieved the respective PFAM domains [34] from SIMAP [35]. We selected protein domains for which abundance significantly differed between these organisms by applying a Fisher's Exact Test. Table S18 shows the PFAM domains with statistical significance (p-value < 0.01).

Table S19: Over- and underrepresented PFAM domains in the genome of *Sorghum bicolor*

The table depicts the comparison of percentages and absolute numbers of the respective over- and underrepresented domains. With respect to rice, the RAP2 annotation has been used. Fields highlighted in green depict domain signatures that are overrepresented in Sorghum while fields highlighted in red depict domains underrepresented in Sorghum. Some gene families amplified in sorghum contain motifs associated with transposable elements (Zinc knuckle, haT family dimerization domain, reverse transcriptase, transposase family *Tnp2*, transposase DDE domain, and putative gypsy type transposons), and may have escaped repeat masking or evolved from transposable elements by neofunctionalization, the latter consistent with their presence

in the manually annotated rice genome. Protein domains that are indicative of transposable elements are marked with an asterisk.

Pfam		Description	all	Sb	[%]	Os	[%]	At	[%]	Pt	[%]	p-val
PF00067		Cytochrome P450	1160	326	1.10	228	0.98	283	0.79	323	0.86	1.6E-04
PF00098	*	Zinc knuckle	434	179	0.60	64	0.27	91	0.25	100	0.27	2.1E-16
PF01370		NAD dependent epimerase/dehydratase family	520	154	0.52	108	0.46	114	0.32	144	0.38	8.0E-04
PF04434		SWIM zinc finger	347	151	0.51	77	0.33	83	0.23	36	0.10	1.8E-16
PF00651		BTB/POZ domain	291	127	0.43	66	0.28	54	0.15	44	0.12	3.2E-14
PF07993		Male sterility protein	372	109	0.37	79	0.34	90	0.25	94	0.25	5.9E-03
PF05699		hAT family dimerisation domain	281	100	0.34	66	0.28	36	0.10	79	0.21	3.5E-06
PF03101		FAR1 DNA-binding domain	193	91	0.31	39	0.17	26	0.07	37	0.10	6.4E-13
PF03330		Rare lipoprotein A (RlpA)-like double-psi beta-barrel	241	91	0.31	51	0.22	42	0.12	57	0.15	5.3E-07
PF00026		Eukaryotic aspartyl protease	285	86	0.29	56	0.24	73	0.20	70	0.19	5.9E-03
PF00917		MATH domain	232	85	0.29	41	0.18	82	0.23	24	0.06	4.9E-06
PF01357		Pollen allergen	220	82	0.28	58	0.25	40	0.11	40	0.11	3.4E-06
PF03372	*	Endonuclease/Exonuclease/phosphatase family	234	75	0.25	30	0.13	68	0.19	61	0.16	1.8E-03
PF00078	*	Reverse transcriptase (RNA-dependent DNA polymerase)	168	68	0.23	22	0.09	66	0.18	12	0.03	7.8E-07
PF05970		Eukaryotic protein of unknown function (DUF889)	116	66	0.22	18	0.08	30	0.08	2	0.01	1.3E-14
PF04578		Protein of unknown function, DUF594	146	63	0.21	41	0.18	6	0.02	36	0.10	1.3E-07
PF08330		Protein of unknown function (DUF1723)	118	54	0.18	31	0.13	17	0.05	16	0.04	9.8E-08
PF08541		3-Oxoacyl-[acyl-carrier-protein (ACP)] synthase III C terminal	163	52	0.17	38	0.16	25	0.07	48	0.13	9.0E-03
PF02797		Chalcone and stilbene synthases, C-terminal domain	159	51	0.17	36	0.15	25	0.07	47	0.13	8.6E-03
PF07762		Protein of unknown function (DUF1618)	84	50	0.17	34	0.15	0	0.00	0	0.00	1.9E-12
PF00891		O-methyltransferase	134	46	0.15	33	0.14	21	0.06	34	0.09	3.0E-03
PF07893		Protein of unknown function (DUF1668)	80	45	0.15	35	0.15	0	0.00	0	0.00	3.2E-10
PF00248		Aldo/keto reductase family	131	44	0.15	25	0.11	27	0.08	35	0.09	5.7E-03
PF04937		Protein of unknown function (DUF 659)	78	42	0.14	12	0.05	11	0.03	13	0.03	7.4E-09
PF04398		Protein of unknown function, DUF538	126	42	0.14	31	0.13	23	0.06	30	0.08	7.9E-03
PF02992	*	Transposase family tnp2	73	40	0.13	9	0.04	24	0.07	0	0.00	8.6E-09
PF04844		Protein of unknown function, DUF623	101	36	0.12	22	0.09	17	0.05	26	0.07	4.0E-03
PF00195		Chalcone and stilbene synthases, N-terminal domain	73	34	0.11	18	0.08	4	0.01	17	0.05	1.4E-05
PF03087		Arabidopsis protein of unknown function	89	32	0.11	16	0.07	17	0.05	24	0.06	5.6E-03
PF04577		Protein of unknown function (DUF563)	68	29	0.10	22	0.09	9	0.03	8	0.02	3.8E-04
PF01598		NA	74	27	0.09	15	0.06	20	0.06	12	0.03	8.3E-03
PF06839		GRF zinc finger	47	26	0.09	6	0.03	10	0.03	5	0.01	2.7E-06
PF07645		Calcium binding EGF domain	53	26	0.09	18	0.08	5	0.01	4	0.01	4.5E-05
PF00092		von Willebrand factor type A domain	63	24	0.08	14	0.06	15	0.04	10	0.03	6.8E-03
PF00280		Potato inhibitor I family	57	23	0.08	6	0.03	6	0.02	22	0.06	3.5E-03
PF03088		Strictosidine synthase	61	23	0.08	10	0.04	17	0.05	11	0.03	9.2E-03
PF04601		Protein of unknown function (DUF569)	45	21	0.07	8	0.03	7	0.02	9	0.02	5.6E-04
PF01161		Phosphatidylethanolamine-binding protein	48	20	0.07	14	0.06	7	0.02	7	0.02	4.0E-03
PF00079		Serpin (serine protease inhibitor)	39	18	0.06	7	0.03	11	0.03	3	0.01	1.6E-03
PF08450		SMP-30/Gluconolaconase/LRE-like region	42	18	0.06	9	0.04	8	0.02	7	0.02	4.4E-03
PF01559		Zein seed storage protein	14	14	0.05	0	0.00	0	0.00	0	0.00	1.6E-09
PF01609	*	Transposase DDE domain	31	14	0.05	9	0.04	6	0.02	2	0.01	6.6E-03

PF08224		Domain of unknown function (DUF1719)	21	12	0.04	9	0.04	0	0.00	0	0.00	9.5E-04
PF08787		Alginate lyase	19	11	0.04	2	0.01	0	0.00	6	0.02	1.3E-03
PF04195	*	Putative gypsy type transposon	10	10	0.03	0	0.00	0	0.00	0	0.00	5.2E-07
PF02496		ABA/WDS induced protein	14	8	0.03	4	0.02	0	0.00	2	0.01	7.0E-03
PF01657		Domain of unknown function DUF26	315	56	0.19	55	0.24	113	0.32	91	0.24	8.2E-03
PF00295		Glycosyl hydrolases family 28	230	38	0.13	34	0.15	80	0.22	78	0.21	6.1E-03
PF00132	*	Bacterial transferase hexapeptide (three repeats)	126	16	0.05	19	0.08	33	0.09	58	0.15	1.7E-03
PF00407		Pathogenesis-related protein Bet v I family	106	14	0.05	12	0.05	33	0.09	47	0.13	6.0E-03
PF07649		C1-like domain	223	7	0.02	7	0.03	176	0.49	33	0.09	1.5E-17
PF03107		C1 domain	201	5	0.02	4	0.02	166	0.46	26	0.07	2.9E-17
PF04396		Protein of unknown function, DUF537	49	4	0.01	5	0.02	31	0.09	9	0.02	5.0E-03
PF08137		DVL family	48	4	0.01	8	0.03	24	0.07	12	0.03	6.1E-03
PF08268		F-box associated	125	1	0.00	1	0.00	117	0.33	6	0.02	1.1E-13
PF00197		Trypsin and protease inhibitor	30	1	0.00	1	0.00	8	0.02	20	0.05	3.3E-03
PF08491		Squalene epoxidase	26	1	0.00	2	0.01	7	0.02	16	0.04	8.4E-03
PF07734		F-box associated	181	0	0.00	1	0.00	160	0.45	20	0.05	8.0E-22

S4.7 Protein family comparison across angiosperms

To identify and estimate the size of gene families in the *Sorghum* genome we applied OrthoMCL [36]. OrthoMCL allows one to infer potentially-orthologous groups of proteins across *Sorghum bicolor, Arabidopsis thaliana, Oryza sativa and Populus trichocarpa*. It can group orthologous as well as paralogous sequences over multiple eukaryotic taxa by using a Markov Cluster algorithm [37]. The algorithm calculates global rather than local similarities to cluster proteins into families. Consequently, proteins are not clustered according to individual protein domains but to overall conservation. We used the OrthoMCL standard settings (Blastp e-value < 1e-05) to compute the all-against-all similarities. See Figure 4 in the main text.

S4.8 Sorghum specific protein families

In addition to the interspecies comparison of protein families, we compared the PFAM domains of proteins families that are specific for sorghum with the PFAM domains of the protein families that are shared with *Arabidopsis*, rice and poplar, respectively (Table S20).

Table S20: Over- and underrepresented PFAM domains of *Sorghum* specific protein families

The table depicts the comparison of percentages and absolute numbers of the respective over- and underrepresented domains. Fields highlighted in green represent PFAM domains that are overrepresented in sorghum specific protein families while fields highlighted in red represent underrepresented domains. Protein domains which are indicative of transposable elements are marked with an asterisk.

Pfam-Domain		Description	all	Sb specific OrthoMCL	[%]	Sb non-specific OrthoMCL	[%]	p-value
PF00646		F-box domain	442	183	9.69	259	1.08	2.2E-91
PF00098	*	Zinc knuckle	161	51	2.70	110	0.46	6.0E-20
PF00097		Zinc finger, C3HC4 type (RING finger)	298	39	2.07	259	1.08	2.9E-04
PF00651		BTB/POZ domain	86	38	2.01	48	0.20	5.2E-21
PF04434		SWIM zinc finger	142	36	1.91	106	0.44	2.6E-11
PF02992	*	Transposase family tnp2	37	35	1.85	2	0.01	7.0E-38
PF04578		Protein of unknown function, DUF594	46	32	1.69	14	0.06	2.8E-26
PF00249		Myb-like DNA-binding domain	213	27	1.43	186	0.78	3.6E-03
PF00931		NB-ARC domain	198	26	1.38	172	0.72	2.6E-03
PF08330		Protein of unknown function (DUF1723)	45	25	1.32	20	0.08	2.5E-17
PF07723		Leucine Rich Repeat	77	25	1.32	52	0.22	9.1E-11
PF00917		MATH domain	60	24	1.27	36	0.15	1.2E-12
PF02902		Ulp1 protease family, C-terminal catalytic domain	28	21	1.11	7	0.03	8.6E-19
PF00106		short chain dehydrogenase	125	21	1.11	104	0.43	2.7E-04
PF01370		NAD dependent epimerase/dehydratase family	134	20	1.06	114	0.48	1.7E-03
PF03478		Protein of unknown function (DUF295)	58	19	1.01	39	0.16	1.4E-08
PF00026		Eukaryotic aspartyl protease	71	19	1.01	52	0.22	5.0E-07
PF00234		Protease inhibitor/seed storage/LTP family	88	19	1.01	69	0.29	1.6E-05
PF07893		Protein of unknown function (DUF1668)	38	18	0.95	20	0.08	2.6E-11
PF03101		FAR1 DNA-binding domain	86	18	0.95	68	0.28	4.0E-05
PF00891		O-methyltransferase	37	17	0.90	20	0.08	1.7E-10
PF00023		Ankyrin repeat	113	17	0.90	96	0.40	3.4E-03
PF00141		Peroxidase	120	17	0.90	103	0.43	6.3E-03
PF03330		Rare lipoprotein A (RlpA)-like double-psi beta-barrel	78	16	0.85	62	0.26	1.4E-04
PF07993		Male sterility protein	96	16	0.85	80	0.33	1.5E-03
PF08387		FBD	30	15	0.79	15	0.06	4.5E-10
PF00657		GDSL-like Lipase/Acylhydrolase	106	15	0.79	91	0.38	1.0E-02
PF01559		Zein seed storage protein	14	14	0.74	0	0.00	1.2E-16
PF02519		Auxin responsive protein	66	14	0.74	52	0.22	2.4E-04
PF02362		B3 DNA binding domain	71	14	0.74	57	0.24	5.4E-04
PF08659		KR domain	58	12	0.64	46	0.19	8.4E-04
PF05699		hAT family dimerisation domain	76	12	0.64	64	0.27	8.7E-03
PF01357		Pollen allergen	68	11	0.58	57	0.24	9.8E-03
PF04195	*	Putative gypsy type transposon	10	10	0.53	0	0.00	4.2E-12
PF06839		GRF zinc finger	25	10	0.53	15	0.06	4.9E-06
PF03936		Terpene synthase family, metal binding domain	27	10	0.53	17	0.07	1.1E-05
PF08100		Dimerisation domain	29	10	0.53	19	0.08	2.3E-05
PF08246		Cathepsin propeptide inhibitor domain (I29)	38	10	0.53	28	0.12	3.0E-04
PF08787		Alginate lyase	11	9	0.48	2	0.01	2.8E-09
PF00079		Serpin (serine protease inhibitor)	15	9	0.48	6	0.03	1.9E-07
PF00112		Papain family cysteine protease	38	9	0.48	29	0.12	1.3E-03
PF01609	*	Transposase DDE domain	11	8	0.42	3	0.01	1.1E-07
PF00092		von Willebrand factor type A domain	21	8	0.42	13	0.05	6.8E-05
PF01598		NA	21	8	0.42	13	0.05	6.8E-05
PF01397		Terpene synthase, N-terminal domain	24	8	0.42	16	0.07	2.0E-04

PF07762		Protein of unknown function (DUF1618)	41	8	0.42	33	0.14	8.6E-03
PF02466		Tim17/Tim22/Tim23 family	19	7	0.37	12	0.05	2.5E-04
PF00665	*	Integrase core domain	30	7	0.37	23	0.10	5.0E-03
PF00314		Thaumatin family	32	7	0.37	25	0.10	7.3E-03
PF01612		3'-5' exonuclease	20	6	0.32	14	0.06	2.4E-03
PF01485		IBR domain	22	6	0.32	16	0.07	4.0E-03
PF00264		Common central domain of tyrosinase	8	5	0.26	3	0.01	9.6E-05
PF09265		Cytokinin dehydrogenase 1, FAD and cytokinin binding	10	5	0.26	5	0.02	3.8E-04
PF03181		BURP domain	11	5	0.26	6	0.03	6.5E-04
PF05241		Emopamil binding protein	6	4	0.21	2	0.01	3.8E-04
PF00187		Chitin recognition protein	8	4	0.21	4	0.02	1.6E-03
PF03405		Fatty acid desaturase	10	4	0.21	6	0.03	4.2E-03
PF03061		Thioesterase superfamily	12	4	0.21	8	0.03	8.7E-03
PF00182		Chitinase class I	12	4	0.21	8	0.03	8.7E-03
PF05129		Transcription elongation factor Elf1 like	4	3	0.16	1	0.00	1.5E-03
PF01255		Putative undecaprenyl diphosphate synthase	5	3	0.16	2	0.01	3.5E-03
PF00304		Gamma-thionin family	6	3	0.16	3	0.01	6.6E-03
PF00069		Protein kinase domain	1052	46	2.44	1006	4.19	4.7E-05
PF07714		Protein tyrosine kinase	596	19	1.01	577	2.41	1.2E-05
PF08263		Leucine rich repeat N-terminal domain	248	6	0.32	242	1.01	6.8E-04
PF01535		PPR repeat	433	6	0.32	427	1.78	1.2E-08
PF00149		Calcineurin-like phosphoesterase	64	0	0.00	64	0.27	7.8E-03
PF02985		HEAT repeat	65	0	0.00	65	0.27	7.2E-03
PF00612		IQ calmodulin-binding motif	66	0	0.00	66	0.28	6.7E-03
PF08241		Methyltransferase domain	67	0	0.00	67	0.28	6.2E-03
PF03106		WRKY DNA -binding domain	68	0	0.00	68	0.28	5.7E-03
PF00168		C2 domain	74	0	0.00	74	0.31	3.6E-03
PF00702		haloacid dehalogenase-like hydrolase	76	0	0.00	76	0.32	3.1E-03
PF00270		DEAD/DEAH box helicase	77	0	0.00	77	0.32	2.9E-03
PF07719		Tetratricopeptide repeat	95	0	0.00	95	0.40	7.4E-04
PF00515		Tetratricopeptide repeat	97	0	0.00	97	0.40	6.3E-04
PF00226		DnaJ domain	101	0	0.00	101	0.42	4.7E-04
PF00271		Helicase conserved C-terminal domain	124	0	0.00	124	0.52	8.1E-05
PF00005		ABC transporter	128	0	0.00	128	0.53	6.0E-05
PF00400		WD domain, G-beta repeat	182	0	0.00	182	0.76	9.8E-07

S5. Gene structure and comparison with rice

The distribution of coding exon lengths is peaked at 80 bp, with median coding exon length 140 bp. The distribution of coding exon lengths is essentially the same between rice and sorghum (Figure S6). The percent identity between rice and sorghum coding sequences is peaked at 85% and the distribution trails off at 75% (Figure S6).

Figure S6. Coding exon length distributions for sorghum (red) and rice (green)

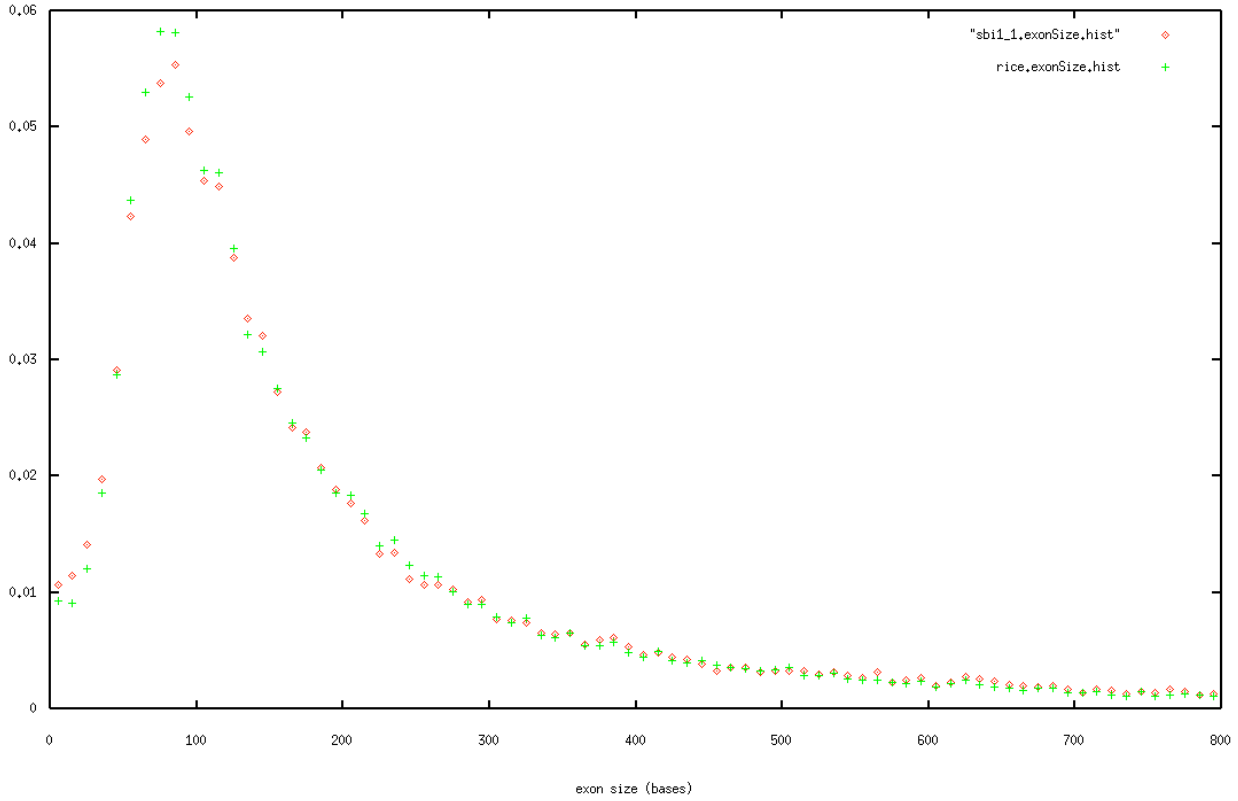
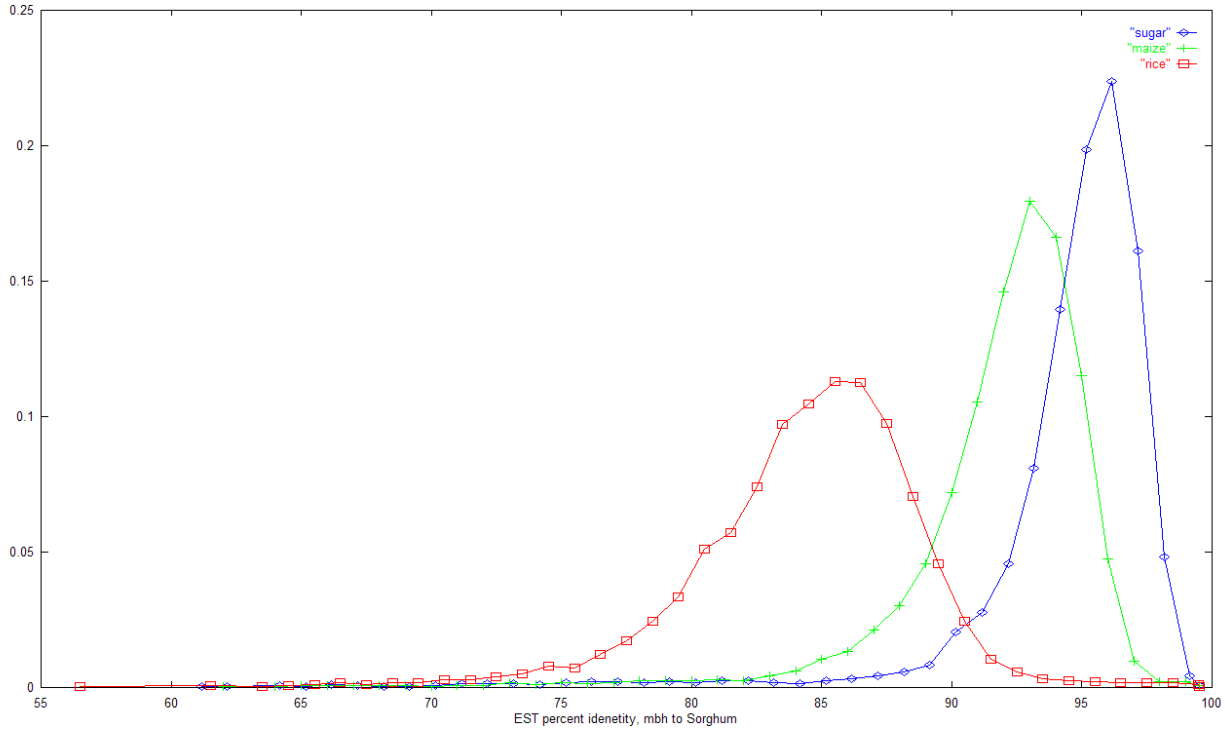


Table S21: Statistics of sorghum gene composition.

Features	OS (Rap2)	ZM*	SB (1.3)
Size	372,089,805.0	14,380,000.0	738,540,932.0
Chromosome assemblies			659,229,367.0
unassembled			79,311,565.0
# genes	29,389.0	330.0	27,458.0
# exons		1,520.0	129,411.0
average # of exons per gene		4.6	4.7
average exon size [bp]		259.0	268.0
median exon size [bp]			139.0
average intron size [bp]	440.0	607.0	436.0
median intron size [bp]	161.0		144.0
average gene size [bp] with UTR	3,340.0		
median gene size [bp] with UTR	2,734.0		
average gene size [bp] (without UTR)			2,873.0
median gene size [bp] (without UTR)			2,116.0
Average gene density (kb per gene)	12.7	43.6	24.0

100 random BACs*

Figure S7. Nucleotide identity between sorghum transcripts and sugarcane (blue), maize (green), and rice (red) (based on assembled ESTs)



S5.4 CISP identification.

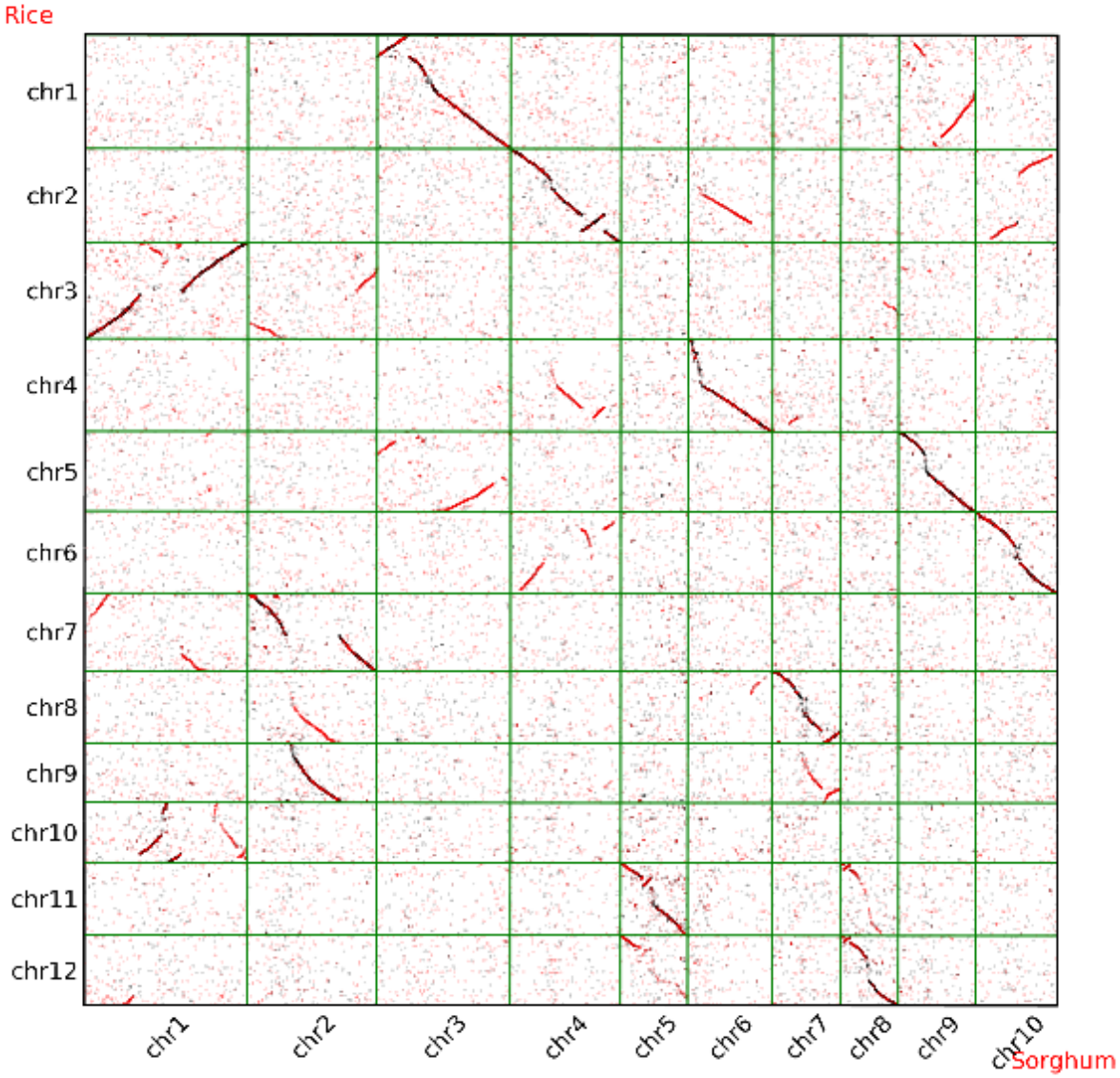
For identification of conserved intron-scanning primers, TIGR rice cDNA models (66,710; version 5) were downloaded (ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs /pseudomolecules/version_5.0/all.chrs/) and sorghum contigs (3,294 super; 10 chromosomes) were aligned to rice cDNA using NCBI BLAST 2.2.13 with an e-value cutoff of 1e-50. This resulted in 76,327 hits. Of these alignments, 6,760 unique rice cDNAs hit the sorghum genome that had at least one pair of exons with perfectly conserved HSP fragments that were at least 20bp in length and no more than 2000bp apart. A total of 3,694 of these genes had a single exon pair. These exon pairs were extracted from the BLAST report with cisp_extractor.pl (v1.2; F. A. Feltus, author), and provided in Supplementary List #2.

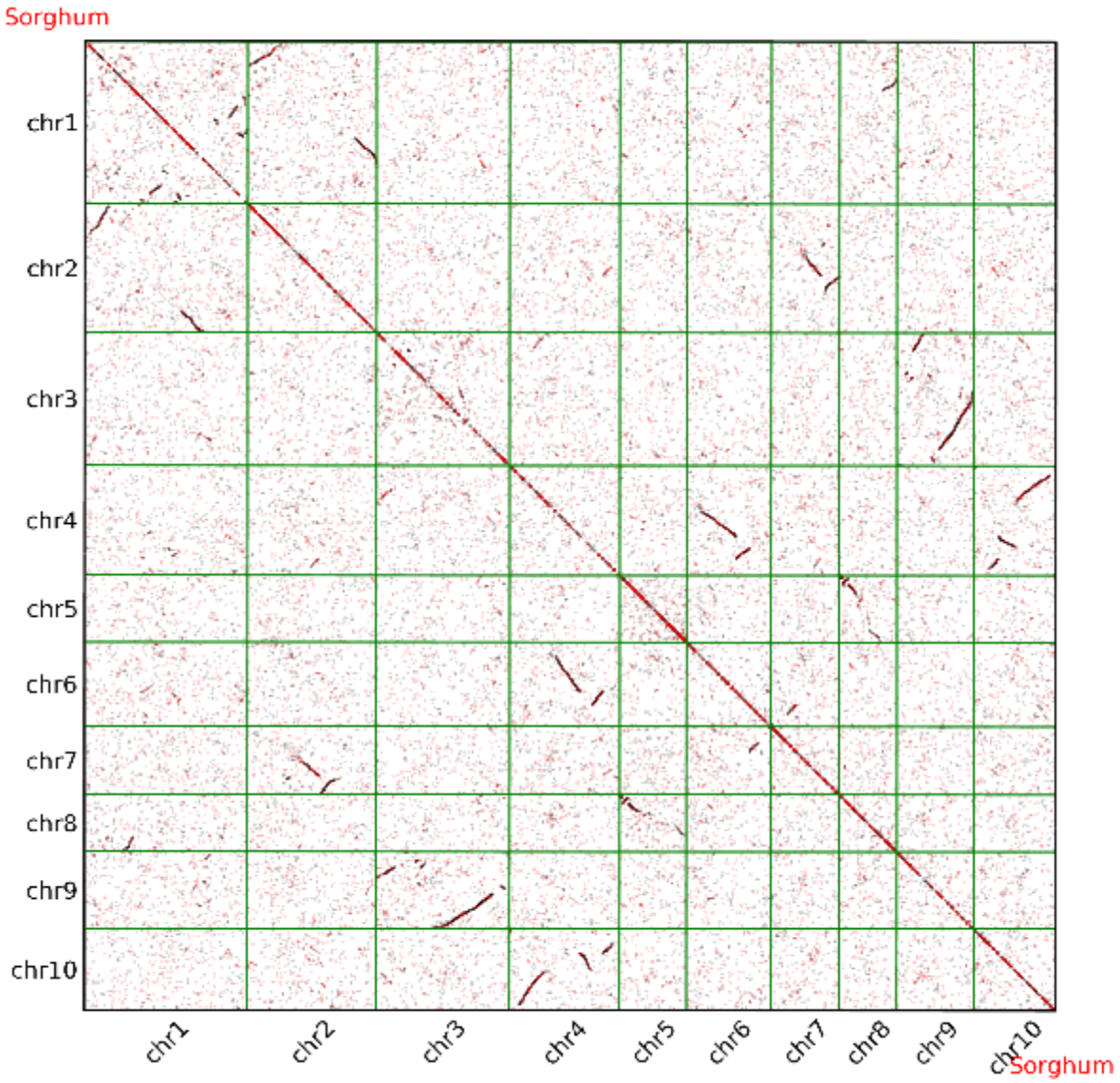
Supplemental Note 6. Identification and characterization of segments of conserved synteny.

Comparative dot plots used ColinearScan [4] and multi-alignments used MCScan [5], applied to rice annotation project (RAP) data version 2 (mapped representative models, 29389 loci) and *Sorghum bicolor* sbi1.4 annotation set (34496 loci). We then did

pairwise BLASTP (*E* < 1e-5, top five hits), both within each genome and between the two genomes to retrieve potential anchors. *Zea* BAC sequences and FPC contig coordinates on a virtual *Zea* genome were downloaded from the Maize Genome Browser (http://www.maizesequence.org, release Jan. 7, 2008). *Sorghum* coding sequences were then searched against *Zea* BACs for potential orthologous *Zea* genes using translated BLAT [6] with minimum score 100.

Figure S8. Global dot-plot of *Oryza* - *Sorghum* and *Sorghum* – *Sorghum* genome alignments. Global dot-plot between Oryza and Sorghum and within Sorghum. Scales are gene indices on the chromosomes. We did not use basepair scales since most synteny lie within relatively compact recombinogenic regions. For each graph, x-axis is the query, y-axis is the database for BLASTP and top two hits were plotted. Black dots are best hits, red dots are second best hits.





S6.1 Pfam domains enriched in singleton or syntenic duplicated genes of sorghum.

We have previously detailed non-random patterns in the retention or loss of gene following the ancient genome duplication shared by rice and sorghum, based on the rice sequence. Analysis of the sorghum sequence using the methods described [38] largely corroborates earlier findings, with all duplication-enriched rice gene families also duplication-enriched in sorghum, and most singleton-enriched gene families also singleton-enriched in sorghum (although in a few cases, sorghum has too few such genes for the observation to be statistically significant). A total of 2 novel duplication-resistant gene families and 10 novel singleton-enriched gene families were found in

sorghum, based on the stringent (0.001) statistical criteria that we have applied previously, and are listed below.

pfam	Number of Syntenic duplicates	Number of singletons	Deviation from random	Sorghum class
PF07714	207	2	2.9E-12	duplicate enriched
PF08263	80	2	5.3E-05	duplicate enriched
PF02985	7	16	5.1E-09	singleton enriched
PF00098	4	10	2.2E-06	singleton enriched
PF08242	4	9	1.3E-05	singleton enriched
PF00288	2	7	2.0E-05	singleton enriched
PF08544	2	6	1.3E-04	singleton enriched
PF01494		4	4.3E-04	singleton enriched
PF01979		4	4.3E-04	singleton enriched
PF03810		4	4.3E-04	singleton enriched
PF06747		4	4.3 E-04	singleton enriched
PF00300	1	4	9.6E-04	singleton enriched

Supplemental Note 7. Timing and characterization of grass-specific genome duplication.

Counting transversions at four-fold synonymous sites, and correcting for multiple transversions (averaging over the long syntenic blocks to get good signal) we find:

Sorghum genome duplication:	4DTv-obs=0.315	=>	4DTv-corr=0.497
Rice genome duplication:	4DTv-obs=0.28	=>	4DTv-corr=0.411
Sorghum-rice divergence:	4DTv-obs=0.24	=>	4DTv-corr=0.327

Clearly the genome duplication is more ancient than the speciation, consistent with it being shared [20]. However, interestingly, sorghum appears to be acquiring more substitutions/site than rice since the duplication (0.315>0.28).

We can use these three (corrected for multiple hits) numbers and computed branch lengths on a tree allowing for independent rates of evolution, with
a = rice-sorghum progenitor transversions/site prior to speciation but after duplication.
b = rice transversions/site since rice-sorghum speciation.
c = sorghum transversions/site since rice-sorghum speciation.

These numbers work out to be:
a = 0.064 transversions/site.
b = 0.142 transversions/site.
c = 0.185 transversions/site.

If we average the rice and sorghum rates (b+c)/2 ~ 0.163, and use a (fossil-based)

estimate for that speciation at 50 million years, then the time from the duplication to the speciation is ~(0.064/0.163)*50 My ~ 20 My which is in keeping with the often-quoted "70 Mya" date [20] for the duplication.

Supplemental Note 8. DNA alignments

We used the VISTA pipeline infrastructure [39] for the construction of genome-wide pairwise DNA alignments between *Sorghum*, the assembly of the Rice v.5.0 genome and 9527 Maize BACs from Genbank (retrieved on June 1, 2007; total length - 1.55Gbp). To align genomes we have implemented algorithms that used an efficient combination of global and local alignment methods. First, we obtained a map of large blocks of conserved synteny between the two species by applying Shuffle-LAGAN glocal chaining algorithm [40] to local alignments produced by translated BLAT [16]. After that we used Supermap, the fully symmetric whole-genome extension to the Shuffle-LAGAN. Then, in each syntenic block we applied Shuffle-LAGAN a second time to obtain a more fine-grained map of small-scale rearrangements such as inversions.

Coverage of different functional intervals of the sorghum genomes by alignment (Table S21) was calculated using the technique first applied to the human-mouse comparison [41]). Both sorghum-rice and sorghum-maize alignments demonstrate high level of DNA conservation between species. 39.9% of all aligned to rice sorghum sequence are conserved at the 70%/100bp level (65% for the maize alignment). A total of 77.5% of the length of sorghum exons are covered by the alignment with rice and 87.3% of base pairs in these exon alignments belong to intervals with a high level of conservation (above 70%/100 bp). These numbers for the maize alignment are equal 63.3 and 92% accordingly. Aligned non-coding regions of sorghum contain about 12% of highly conserved with rice intervals, and this percentage is especially high for the alignment with maize – 56%. These intervals can be either under predicted by current techniques coding regions, or other functional elements.

Table S21: Coverage of different intervals of the sorghum genome by the alignments with the rice v.5.0 genome and 1.55Gbp of Maize BACs.

	Rice	Maize
Total coverage:	11.3%	13.52%
utr coverage:	52.7%	60.31%
exons coverage:	77.54%	63.26%
up100 coverage:	38.05%	43.93%
up200 coverage:	33.41%	40.31%
up500 coverage:	25.03%	32.2%
down200 coverage:	28.14%	37.65%

The constructed genome-wide pair-wise alignments can be downloaded from http://pipeline.lbl.gov/downloads.shtml and are accessible for browsing and various types of analysis through the VISTA browser at http://pipeline.lbl.gov/ linked to the JGI *Sorghum* browser.

Figure S9: Multiple VISTA conservation tracks among syntenic regions of plants. Evolutionary relationships among these regions are shown (at right), with black circles representing the pan-cereal duplication (ρ) and a maize-specific duplication (*m*). The region from sorghum chromosome 6: 56.17-56.45Mb is used as the 'reference' (show at top) in the VISTA pipeline[49]. By aligning syntenic regions to sorghum, cases of sub-functionalization in maize can be more easily identified (two are shown).

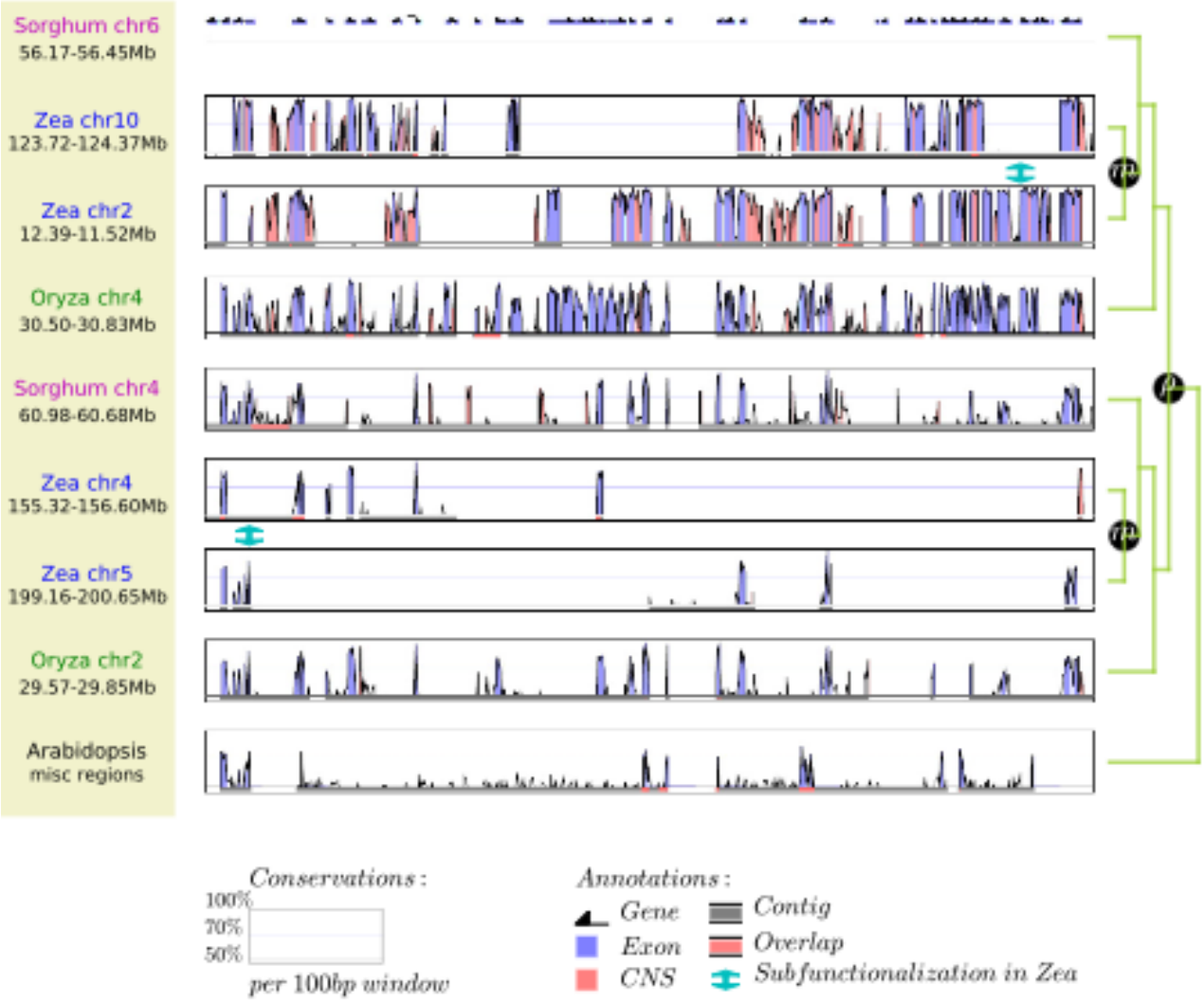
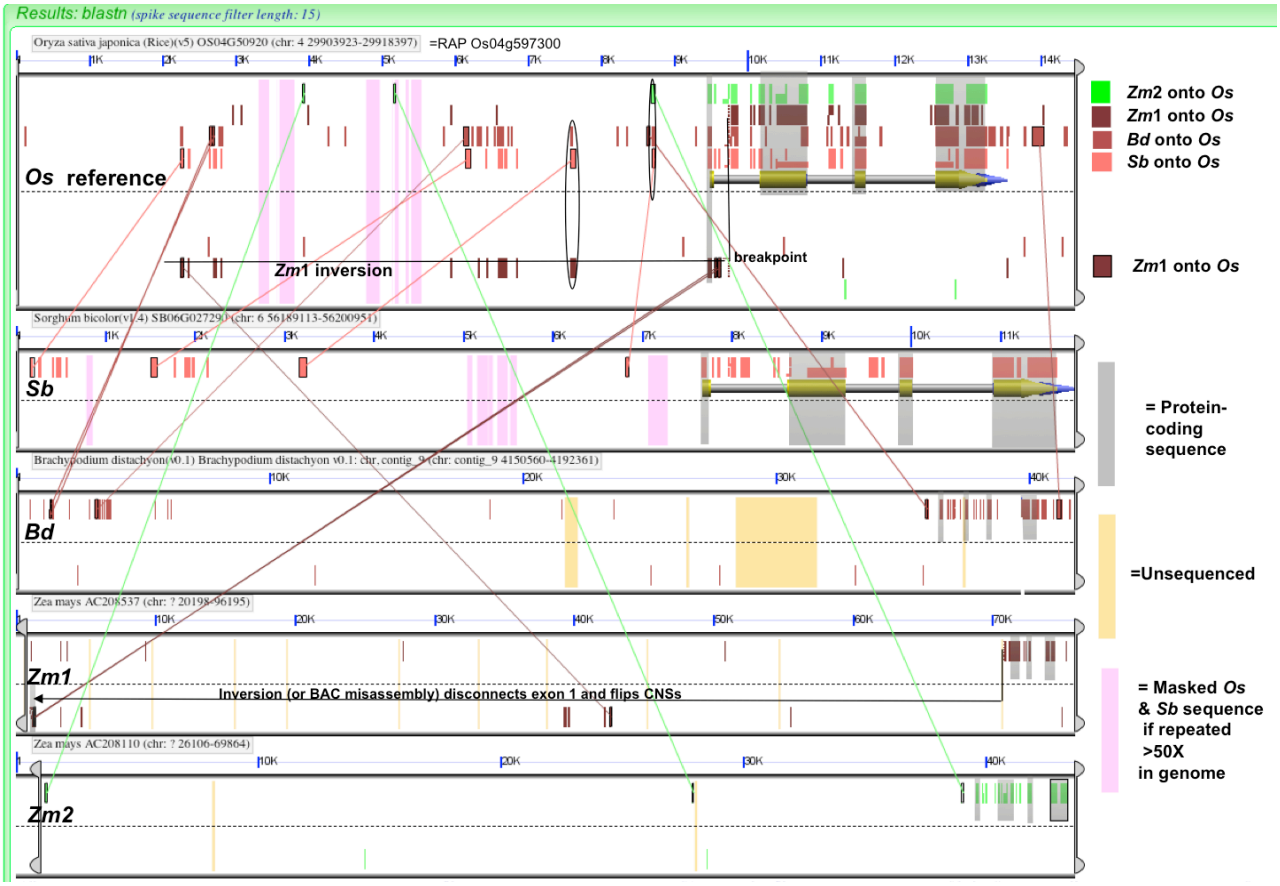


Figure S10: Grass conserved noncoding sequences (CNSs) are often far removed from the genes with which they associate. CNSs are depicted as color-coded boxes—blastn high scoring pairs -- far upstream of a grass *WRKY* transcription factor gene. The rice gene *Os04G50920* (*Os*, upper panel) was used as query against orthologous subject gene regions: sorghum (*Sb*, second panel); *Brachypodium distachyon* (*Bd*, JGIv1, third panel); and the two homeologous maize BACs (*Zm*1 and *Zm*2, AC208537 and 110; maizesequence.org). The exact pairs aligned are color-coded. Masked sequence is also color-coded; without the rice repeat-mask, transposons would obscure this graphic. The four exons are each encased in grey rectangles. Most nonexon hits to *Os* are syntenic, noncoding, and fit the criteria for plant CNSs[50]. Panel 1 plots all blast hits over the rice gene and 9.5 kb of 5' chromosome. Note how putative CNSs pile-up in the introns and upstream space. For example, ovals enclose two examples where all grass sequences except one maize gene share the same CNS, possibly indicating subfunctionalization following the maize tetraploidy. Lines connect the same CNSs; only a few are drawn to reduce clutter. These blastn hits are generally syntenic, although the *Zm1* BAC has a single "inversion" from the leftmost border to just inside exon 1, which flips the entire region (perhaps indicating BAC misassembly). CNSs that are ~7kb upstream in rice and sorghum are 30kb upstream in *Brachypodium* and even more distant in maize. The sorghum sequence may help to solve the mystery of why noncoding sequences are conserved even so far from exons. To regenerate this experiment and to change distances, algorithm, or settings click http://tinyurl.com/6jk5dd.



Supplemental Note 9. Evolution of C4 photosynthesis genes

We identified 7 C4 photosynthesis enzyme genes in the sorghum genome, including 1 phosphoenolpyruvate carboxylase gene (pepc), 1 phosphoenolpyruvate carboxylase kinase gene (ppck), 1 pyruvate orthophosphate dikinase gene (ppdk), 2 carbonic anhydrase genes (cah), 1 malate dehydrogenase gene (mdh), and 1 malic enzyme gene (me).

Known photosynthesis genes in sorghum and maize (Table S22a) were downloaded from the NCBI CoreNucleotide database (http://www.ncbi.nlm.nih.gov/). By searching these known genes against sorghum and rice gene models by running BLAST and by constructing gene trees, the sorghum C4 genes and their isoforms were identified. Neighbor-joining topologies (Figure S9) were generated as the consensus of 100 bootstrap alignment replicates by running MEGA [42]. By searching for gene colinearity in duplicated regions in rice and sorghum genomes using MCSCAN [43], we identified those rice-sorghum orthologs that had preserved gene colinearity (Figure S9).

The C4 pepc gene Sb10g021330.1 shares ~99% amino acid similarity to the one previously identified copy in Sorghum bicolor (GenBank accession no: X17379), and ~93% to the maize C4 gene (NM_001111948) [44]. The corresponding rice ortholog has been lost (Table S22b). Sb10g021330.1 is suggested to be the C4 enzyme gene in that it shares 98.5% sequence identity with cDNA clone HHU2, for which transcripts accumulated more than 20 times higher in mesophyll than in bundle-sheath cells [45]

The likely Sorghum C4 ppck gene Sb04g036570, sharing 99.8% similarity to GenBank item DQ386731, is grouped together with the maize C4 ppck gene (NM_001112338)[46] sharing ~93% amino acid identity.

The Sorghum C4 ppdk gene Sb09g019930 shares ~93% amino acid identity with its maize ortholog (NM_001112268). They share a single rice ortholog Os05g0405000. Sb09g019930 is suggested to be the C4 enzyme gene in that it shares ~95% sequence identity with cDNA clone HHU1, for which transcripts accumulated more than 10-20 times higher in mesophyll than in bundle-sheath cells [45].

There are two sorghum mdh genes, Sb07g023910 (GenBank accession no: S55884) and Sb07g023920 (X53453), which are in tandem locations. However, a previous report indicated that only the latter is light induced, being possibly involved in the C4 pathway [47]. They share a single maize C4 ortholog (X16084) and single non-C4 rice ortholog (Os08g0562100).

The likely sorghum C4 me gene Sb03g003230 (AY274836) shares ~95% amino acid identity with a maize C4 gene (NM_001111843) [48]. There is a tandem me gene copy in sorghum (Sb03g003220), which was likely produced before sorghum-maize divergence (Figure S9). Sb03g003230 is inferred to be involved in C4 pathway based on its 100% identity to cDNA clone HHU3, for which transcripts accumulated in bundle-sheath cells 20 times more than those in mesophyll cells [45]. Comparatively, Sb03g003220 has 95% identity to HHU3, and matches no other cDNA isolated.

There are two types of cah genes: alpha and beta types, with the two gene types sharing relatively low sequence similarity. The C4 cah genes Sb03g029170 and Sb03g029180 are beta-type, and were identified based on their similarity to previously reported clones [45]. Clone HHU69 is ~100% identical to the terminal region of Sb03g029170 and HHU68 ~99% identical to the terminal region of Sb03g029180. Both clones are transcribed in mesophyll cells only, suggesting that they are probably from C4 pathway genes. These clones share relatively lower similarity with the other tandem gene sequences. Alpha-type cah genes include Sb07g022860, Sb07g022880, Sb07g022890, Sb07g022910, Sb10g023940, Sb05g003270, Sb06g015600.

Table S22a. Possible C4 genes identified in the sorghum genome

Gene type	GENE ID	CDSLEN	Related Accession	Maize ortholog
carbonic anhydrase	Sb03g029170	1371		NM_001111889
carbonic anhydrase	Sb03g029180	615		NM_001111889
malate dehydrogenase	Sb07g023920	1290	X53453	X16084
malic enzyme	Sb03g003230	1941	AY274836	NM_001111843, NM_001111913
phosphoenolpyruvate carboxylase	Sb10g021330	2886	X17379	NM_001111948
phosphoenolpyruvate carboxylase kinase	Sb04g036570	855	DQ386731	NM_001112338
pyruvate orthophosphate dikinase	Sb09g019930	2847		NM_001112268

Table S22b. Sorghum C4 genes and their isoforms and their corresponding rice orthologs[1]. C4 genes are <u>underlined</u>

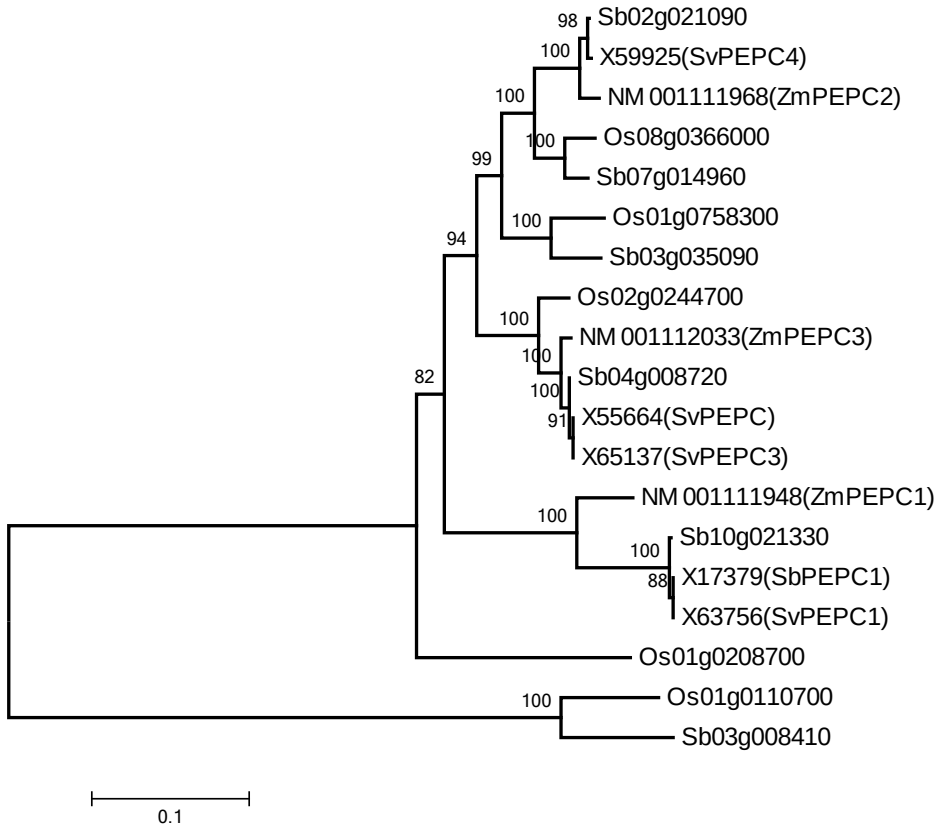
C4 genes and isoforms	Rice gene 1	Sorghum gene 1	Rice gene 2	Sorghum gene 2
carbonic anhydrase	Os01g0639900	<u>Sb03g029170</u>, <u>Sb03g029180</u>, Sb03g029190, Sb03g029200	N.A.	N.A.
malate dehydrogenase	Os08g0562100	N.A.	N.A.	Sb07g023910, <u>Sb07g023920</u>
malic enzyme	Os01g0188400	Sb03g003220, <u>Sb03g003230</u>	Os05g0186300	Sb09g005810
malic enzyme	Os01g0723400	Sb03g033250	N.A.	N.A.
malic enzyme	Os01g0743500	Sb03g034280	N.A.	N.A.
malic enzyme	N.A.	Sb01g017790	Os10g0503500	N.A.
phosphoenolpyruvate carboxylase	Os01g0110700	Sb03g008410	N.A.	N.A.
phosphoenolpyruvate carboxylase	Os01g0758300	Sb03g035090	N.A.	N.A.
phosphoenolpyruvate carboxylase	Os02g0244700	Sb04g008720	N.A.	<u>Sb10g021330</u>
phosphoenolpyruvate carboxylase	Os08g0366000	N.A.	N.A.	Sb07g014960
phosphoenolpyruvate carboxylase	N.A.	Sb02g021090	Os09g0315700[2]	N.A.
phosphoenolpyruvate	Os02g0625300	Sb04g026490	Os04g0517500	Sb06g022690

carboxylase kinase				
phosphoenolpyruvate carboxylase kinase	Os02g0807000	Sb04g036570	N.A.	N.A.
pyruvate orthophosphate dikinase	N.A.	N.A.	Os05g0405000	Sb09g019930

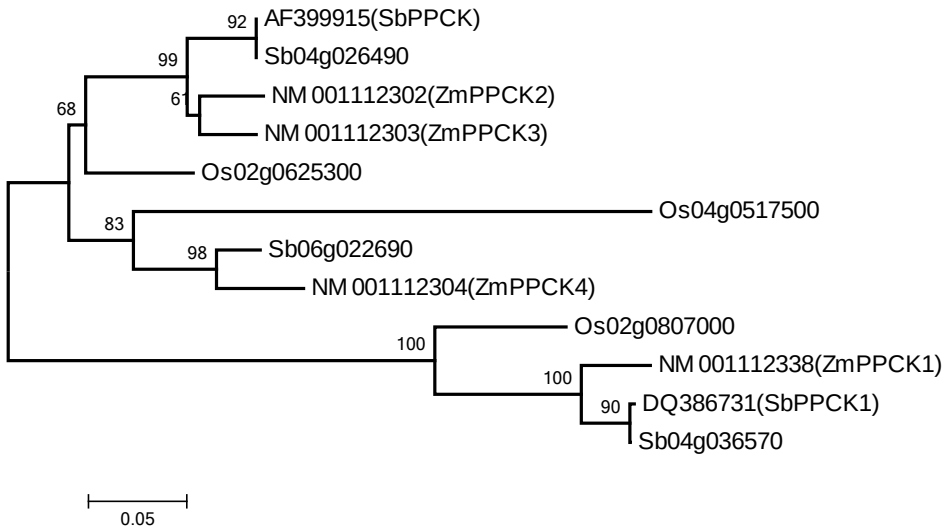
[1]According to gene colinearity, gene1 and gene 2 are paleologs produced by whole genome duplication in the common ancestral genome of rice and sorghum. Rice gene 1 is orthologous to sorghum gene 1 and rice gene 2 is orthologous to sorghum gene2. "N.A." indicates that the anticipated homologous gene is not found at the colinear location, implying possible gene loss or translocation. 2. Os09g0315700 has only partial coding sequence of the other homologs, indicating a possibility that it is a peudo-gene. Therefore, it was not involved in gene tree construction.

Figure S11: Phylogeny of photosynthesis enzyme genes and their isoforms in sorghum, rice and maize. (a) *pepc*; (b) *ppck*; (c) *ppdk*; (d) *cah*; (e) *mdh*; (f) *me*. In the gene ids, "Sb" indicates *Sorghum bicolor* genes, "Sv" indicates *Sorghum vulgare* genes, "Os" indicates *Oryza sativa* genes and "Zm" indicates *Zea mays* genes. Neighbor-joining topologies were generated as the consensus of 100 bootstrap alignment replicates by running MEGA [42].

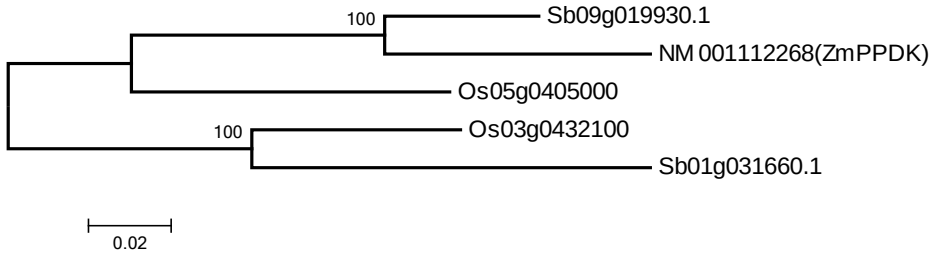
(a). phosphoenolpyruvate carboxylase



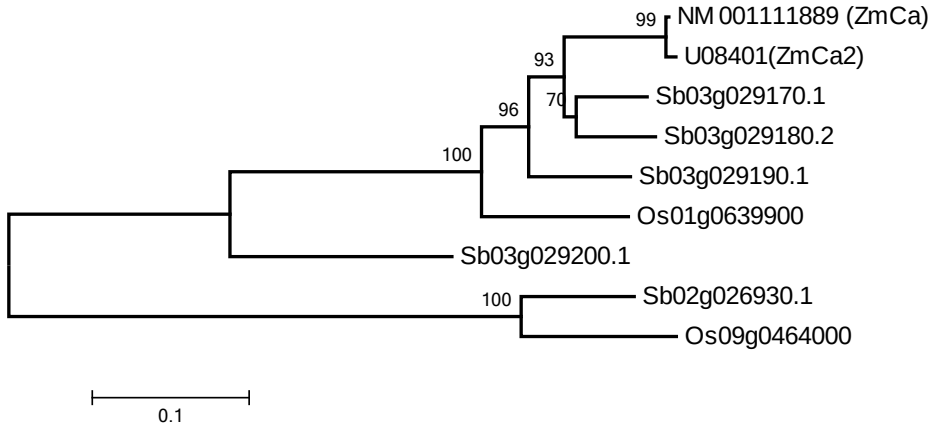
(b). phosphoenolpyruvate carboxylase kinase



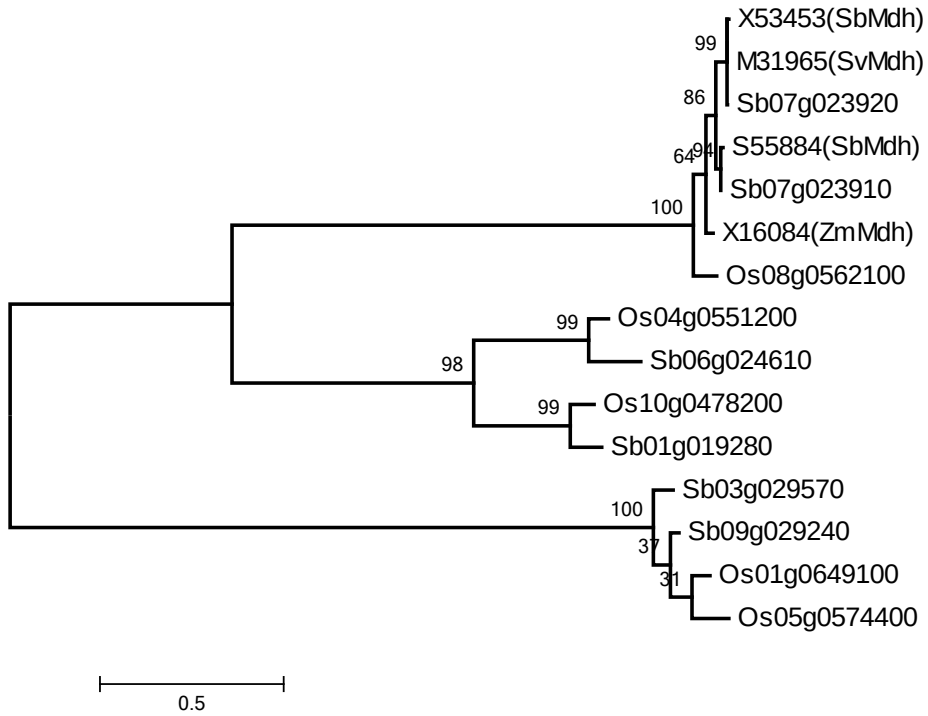
(c). pyruvate orthophosphate dikinase



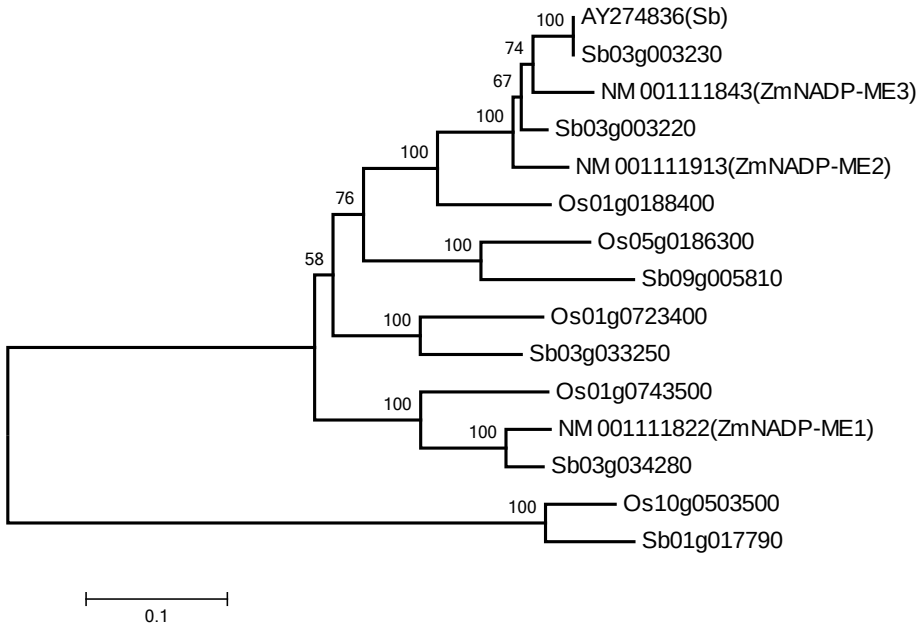
(d). carbonic anhydrase



(e). malate dehydrogenase



(f). malic enzyme



Supplemental Note 10. Evolution of Cell wall synthesis genes

The Carbohydrate-Active Enzyme (CAZy) database (http://www.cazy.org/) contains 91 families of glycosyl transferases (GTs), 112 families of glycosyl hydrolases (GHs), and other carbohydrate-metabolizing enzymes[35].

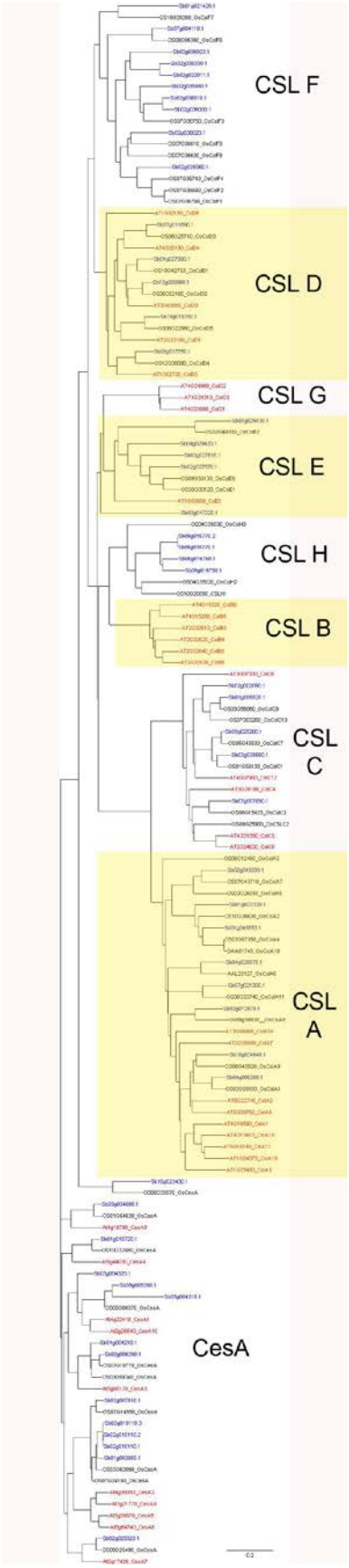
Sorghum cellulose synthase (CesA), cellulose synthase-like (Csl), and a glycosyl transferase, GT31 gene families were constructed by querying the sorghum peptide sequence database, *Sorghum bicolor* sbi1.4 annotation set (34,496 loci), with rice protein sequences from the Purdue cell wall genomics website, http://cellwall.genomics.purdue.edu/, and using NCBI's Basic Local Alignment and Search Tool [1]. A custom DOSshell script was used to direct the BLAST through multiple sequence files using the following parameters: protein-protein BLAST search (BLASTp), expect value of 10^{-20}, and no alignment output. The BLAST results were parsed using a custom C++ script to scan and place the queried rice gene name, associated sorghum gene names, and match score values for any score >200 into a file for later sorting in Microsoft Excel. Duplicate matches due to multiple hits to the same sorghum sequence from closely related rice sequences were eliminated to generate a unique sorghum gene list to extract sorghum gene sequences from the database using the fastacmd program from NCBI [1] in a custom DOSshell script. Table S23 shows relative numbers of family members of *Arabidopsis*, rice, maize, and sorghum broken into families and group clades for Csl and GT31 and potential cell wall expression (primary, secondary, or other/unknown) for CesA.

Dendrograms were assembled from protein coding sequences using the neighbor joining method with ClustalW [49] through the Kyoto University Bioinformatics Center website (http://align.genome.jp/). The parameters used were for a slow, accurate tree with gap open penalty of 10, gap extension penalty of 0.05, and a Gonnet weight matrix for proteins for multiple alignments; a gap open penalty of 10, gap extension penalty of 0.1, and a Gonnet weight matrix for proteins for pairwise alignments. After the initial multiple alignment, individual clade alignments were checked using Multalin [50]. Matches to conserved regions within groups of family clades, were manually checked and non-matching members of the families removed to produce a final tree alignment in ClustalW. Dendrograms were drawn using TreeDyn (http://www.treedyn.org/) [51] and exported as JPEG files. Figure S10 shows the clade structure for the cellulose synthase superfamily, consisting of CesA and Csl sequences. Figure S11 shows the clade structure for family GT31.

Table S23. Comparative numbers of cell-wall genes in families of Arabidopsis, rice, sorghum, and maize

Family #	Family name	Sub Family	Number of genes			
			Arabidopsis	Rice	Sorghum	Maize
2.1	CesA	Primary	3	3	4	5
		Secondary	3	3	3	3
		Other	4	4	5	6
		Total	**10**	**10**	**12**	**14**
2.2	CSL	A	9	11	8	9
		B	6	0	0	0
		C	5	6	5	11
		D	5	5	5	10
		E	1	3	5	4
		F	0	8	10	11
		G	3	0	0	0
		H	0	3	3	1
		Total	**29**	**36**	**36**	**46**
2.3.5	GT31	A	12	7	8	8
		B	6	10	11	10
		C	8	8	8	9
		D	3	2	3	5
		E	3	2	2	3
		F	1	10	6	5
		Total	**33**	**39**	**38**	**40**

Figure S12 (next page). Cellulose synthase superfamily dendrogram for *Arabidopsis*, rice and sorghum. The figure demonstrates good conservation of genes for CesA and Csls between sorghum and rice, with two unique grass clades (CslF and CslH) and two unique *Arabidopsis* clades (CslB and CslG).


CSL F
CSL D
CSL G
CSL E
CSL H
CSL B
CSL C
CSL A
CesA

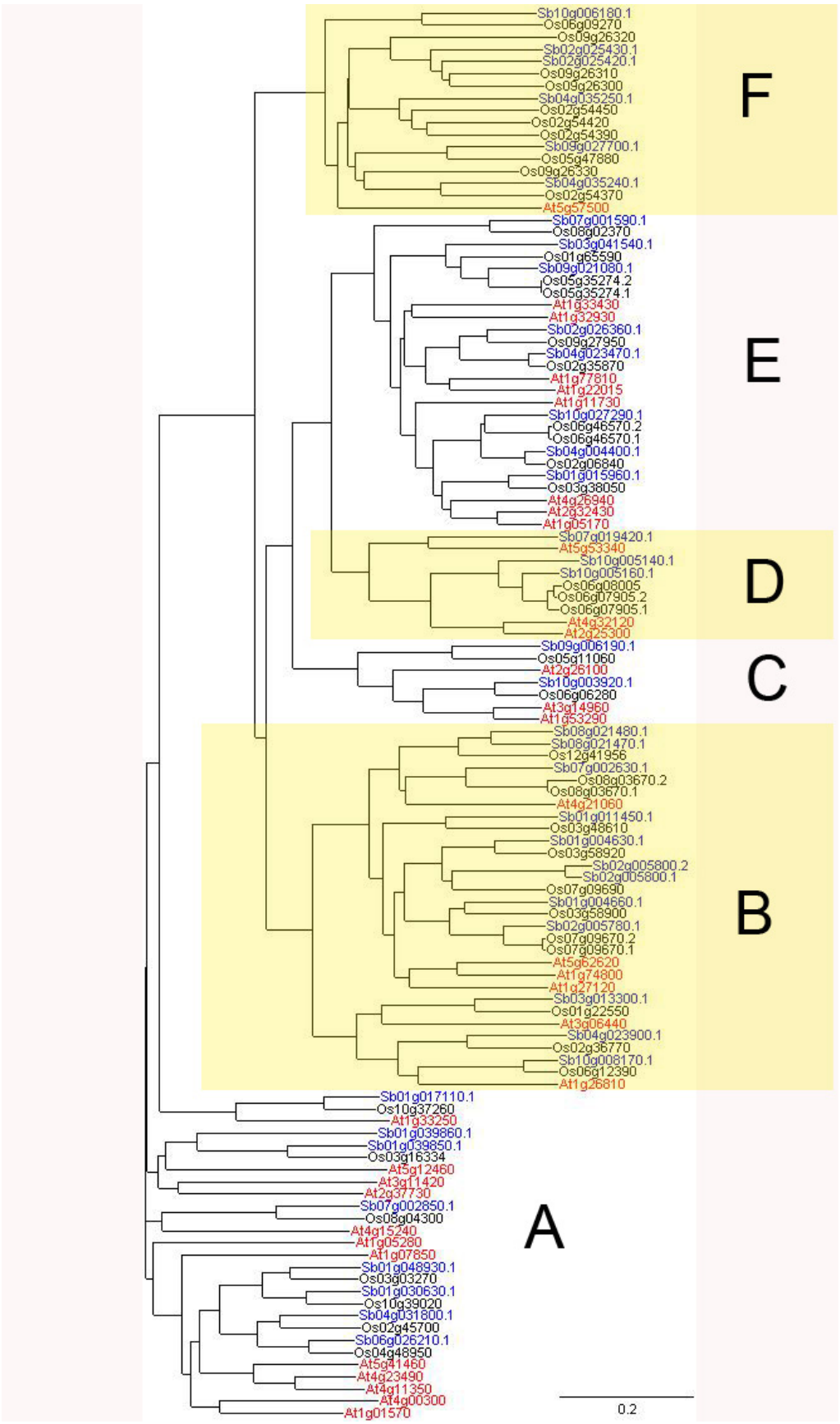

Sb10g006180.1
Os06g09270
Os09g26320
Sb02g025430.1
Sb02g025420.1
Os09g26310
Os09g26300
Sb04g035250.1
Os02g54450
Os02g54420
Os02g54390
Sb09g027700.1
Os05g47880
Os09g26330
Sb04g035240.1
Os02g54370
At5g57500
F
Sb07g001590.1
Os08g02370
Sb03g041540.1
Os01g65590
Sb09g021080.1
Os05g35274.2
Os05g35274.1
At1g33430
At1g32930
Sb02g026360.1
Os09g27950
Sb04g023470.1
Os02g35870
At1g77810
At1g22015
At1g11730
Sb10g027290.1
Os06g46570.2
Os06g46570.1
Sb04g004400.1
Os02g06840
Sb01g015960.1
Os03g38050
At4g26940
At2g32430
At1g05170
E
Sb07g019420.1
At5g53340
Sb10g005140.1
Sb10g005160.1
Os06g08005
Os06g07905.2
Os06g07905.1
At4g32120
At2g25300
D
Sb09g006190.1
Os05g11060
At2g26100
Sb10g003920.1
Os06g06280
At3g14960
At1g53290
C
Sb08g021480.1
Sb08g021470.1
Os12g41956
Sb07g002630.1
Os08g03670.2
Os08g03670.1
At4g21060
Sb01g011450.1
Os03g48610
Sb01g004630.1
Os03g58920
Sb02g005800.2
Sb02g005800.1
Os07g09690
Sb01g004660.1
Os03g58900
Sb02g005780.1
Os07g09670.2
Os07g09670.1
At5g62620
At1g74800
At1g27120
Sb03g013300.1
Os01g22550
At3g06440
Sb04g023900.1
Os02g36770
Sb10g008170.1
Os06g12390
At1g26810
B
Sb01g017110.1
Os10g37260
At1g33250
Sb01g039860.1
Sb01g039850.1
Os03g16334
At5g12460
At3g11420
At2g37730
Sb07g002850.1
Os08g04300
At4g15240
At1g05280
At1g07850
Sb01g048930.1
Os03g03270
Sb01g030630.1
Os10g39020
Sb04g031800.1
Os02g45700
Sb06g026210.1
Os04g48950
At5g41460
At4g23490
At4g11350
At4g00300
At1g01570
A
0.2

Figure S13 (prior page). GT31 dendrogram for Arabidopsis, rice, and sorghum showing clade structure and conservation of genes in the family. The group F genes are greatly expanded in the grasses over Arabidopsis, suggesting possible unique function in grasses. Group A shows a slight reduction in grasses, whereas group B shows a slight expansion.

Supplemental Note 11. Sorghum-sugarcane microcolinearity.

Comparison of sorghum genome to genomic sequences of sugarcane (*Saccharum* spp.) provided insights into the evolutionary history of these closely related diploid and autopolyploid genomes. Twenty selected sugarcane bacterial artificial chromosomes (BACs) were selected for sequencing, two BACs each corresponding to the euchromatic region of individual sorghum chromosome, to study the sequence conservation and synteny. The assembled BAC contigs were annotated using sugarcane and sorghum ESTs. A total of 1.45 Mb sugarcane BAC sequences were unambiguously ordered based on sorghum genome sequence, which accounted for 90% of the estimated 1.6 Mb target BAC sequences. Among the ordered sequences, 986 Kb (68%) collinearly aligned with sorghum sequence. The estimated time of divergence is about 7.7 million years, supporting their recent divergence.

Sugarcane has undergone at least two more rounds of genome-wide duplication to reach its current level of autopolyploidy after its separation with sorghum from a common ancestor. The continuous diploidization of sorghum and the process of polyploidization of sugarcane may result in different gene loss/retention rate. From the aligned sequences, 209 protein coding genes were found in sugarcane, including 155 validated by sugarcane ESTs, 28 by sorghum ESTs, and 26 corresponding to sorghum annotated genes. In homologous region of sorghum, 189 genes were annotated, including 121 validated by sorghum ESTs, 29 by sugarcane ESTs, and 39 from prediction of the most recent version of the annotated sorghum genome. Among these annotated genes, 19 appeared to be sugarcane specific while 12 might be sorghum specific. The larger number of genes from one homolog of the sugarcane genome indicated higher rate of gene loss diploid sorghum genome during it diploidization process. On the other hand, the higher gene retention rate in autopolyploid sugarcane appeared to be against the conventional wisdom of faster gene loss in polypoids because of the existence of a large number of allelic genes.

Among the 20 sequenced sugarcane BACs, 986 kb sequence aligned co-linearly to 1,189 kb sorghum sequence (Figure S7). Aligning homologous genes demonstrated that tandem duplication as a driving force of gene and genome evolution as documented in both sugarcane and sorghum genomes (Figure S13).

Figure S14. Collinear alignment between sugarcane BAC sequences and their sorghum counterparts.

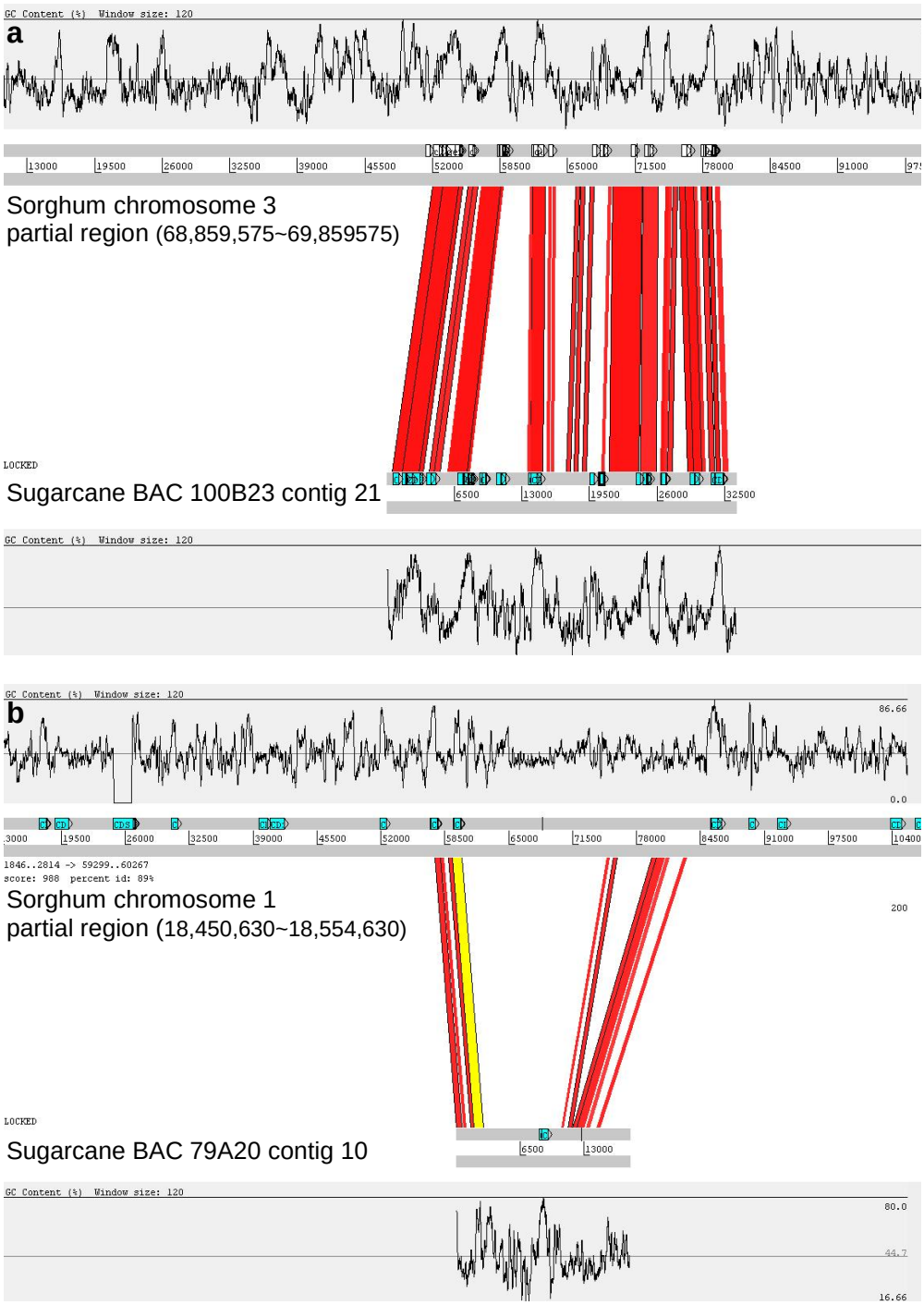
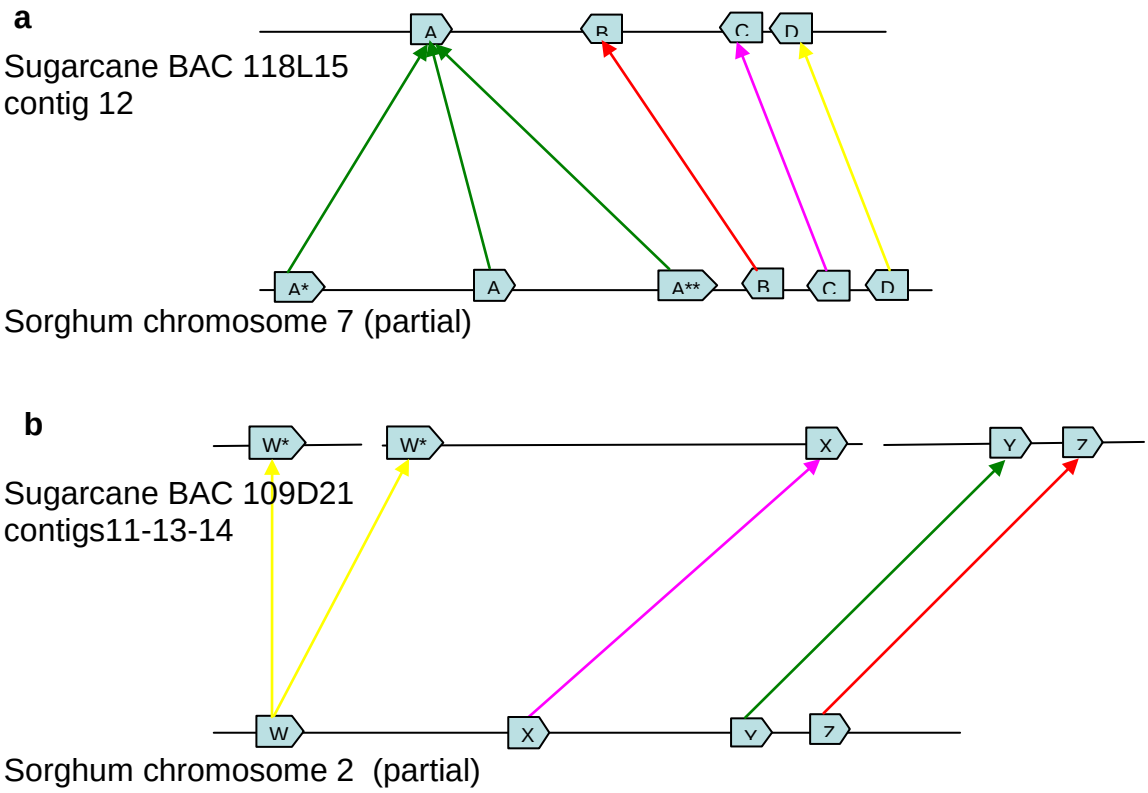


Figure S15. Tandem gene duplication in the sugarcane or sorghum genome. These genes were identified by aligning genome sequence with sorghum ESTs. **a.** Gene duplication on sorghum chromosome 7 but not in the corresponding region of sugarcane BAC 118L15 contig 12. Gene A has three copies in sorghum and only one copy in sugarcane. * indicates the gene missed one exon; ** indicates the gene missed one exon and the other exons are 96% rather than 100% identity to the EST sequence. The putative functions of genes A, B, C, and D are 60S ribosomal protein L10A, expressed protein, fiber protein Fb1, and unknown, respectively. **b.** Gene duplication on sugarcane BAC 109D21 contig 11-13-14 but not in the corresponding region of sorghum chromosome 2. Gene A has two copies in sugarcane and one in sorghum. The putative function of genes A, B, C, and D are serine carboxypeptidase 2, receptor kinase, OSH15 protein, and homeobox transcription factor GNARLY1, respectively.



SUPPLEMENTAL REFERENCES

1 SF Altschul, W Gish, W Miller et al., *J Mol Biol* **215**, 403 (1990).
2 R.-F. Yeh, L. P. Lim, and C. Burge, *Genome Research* **11**, 803 (2001).
3 B.J. Haas, Volfovsky N., Town C.D. et al., *Genome Biology* (3), research0029.1 (2002).
4 X.Y. Wang, X.L. Shi, Z. Li et al., *BMC Bioinformatics* **7**, 447 (2006).
5 H. Tang, J. E. Bowers, X. Wang et al., *Science* **320**, 486 (2008).
6 W. J. Kent, *Genome Res* **12** (4), 656 (2002).
7 R. Spangler, *Australian Systematic Botany* **16**, 279 (2003).
8 H. J. Price, S. L. Dillon, G. Hodnett et al., *Annalso of Botany* **95**, 219 (2005); K Arumuganathan and ED Earle, *Plant Mol Biol Rep.* (9), 208 (1991); M. D. Bennett and I. J. Leitch, *http://www.rbgkew.org.uk/cval/homepage.html*. (2003).
9 R. A. Frederiksen and F. R. Miller, 1972.
10 D. G. Peterson, K. S. Boehm, and S. M. Stack, *Plant Molecular Biology Reporter* **15** (2), 148 (1997).
11 D. G. Peterson, S. R. Schulze, E. B. Sciara et al., *Genome Research* **12** (5), 795 (2002).
12 D. B. Jaffe, J. Butler, S. Gnerre et al., *Genome Research* **13** (1), 91 (2003).
13 J. E. Bowers, M. A. Arias, R. Asher et al., *Proceedings of the National Academy of Sciences of the United States of America* **102** (37), 13206 (2005).
14 J. S. Kim, P. E. Klein, R. R. Klein et al., *Genetics* **169** (2), 1169 (2005).
15 F.A. Feltus, G.E. Hart, K.F. Schertz et al., *Theoretical and Applied Genetics* **112**, 1295 (2006).
16 W. J. Kent, *Genome Research* **4**, 656 (2002).
17 J. E. Bowers, C. Abbey, S. Anderson et al., *Genetics* **165**, 367 (2003).
18 S. Chopra, V. Brendel, J. B. Zhang et al., *Proceedings of the National Academy of Sciences of the United States of America* **96** (26), 15330 (1999).
19 J. T. Miller, S. A. Jackson, S. Nasuda et al., *Theoretical and Applied Genetics* **96** (6-7), 832 (1998).
20 A.H. Paterson, J.E. Bowers, and B. A. Chapman, *Proceedings of the National Academy of Sciences of the United States of America* **101**, 9903 (2004).
21 F. Wei, E. Coe, W. Nelson et al., *PLoS Genet* **3** (7), e123 (2007).
22 D. G. Peterson, S. R. Schulze, E. B. Sciara et al., *Genome Res* **12** (5), 795 (2002).
23 C. Du, J. Caronna, L. He et al., *BMc Genomics* **9**, 51 (2008).
24 K. L. Childs, J. P. Hamilton, W. Zhu et al., *Nucleic Acids Research* **35**, D846 (2007).
25 G. Gremme, V. Brendel, M. E. Sparks et al., *Information and Software Technology* **47** (15), 965 (2005).
26 A. Bairoch, *Molecular & Cellular Proteomics* **4** (8), S2 (2005).
27 D. Swarbreck, C. Wilks, P. Lamesch et al., *Nucleic Acids Research* **36**, D1009 (2008).
28 H. W. Mewes, C. Amid, R. Arnold et al., *Nucleic Acids Research* **32**, D41 (2004).
29 T. Tanaka, B. A. Antonio, S. Kikuchi et al., *Nucleic Acids Research* **36**, D1028 (2008).
30 J. E. Allen and S. L. Salzberg, *Bioinformatics* **21** (18), 3596 (2005).
31 S. M. J. Searle, J. Gilbert, V. Iyer et al., *Genome Research* **14** (5), 963 (2004).
32 B. J. Haas, S. L. Salzberg, W. Zhu et al., *Genome Biology* **9** (2008).
33 S. Griffiths-Jones, H. K. Saini, S. van Dongen et al., *Nucleic Acids Research* **36**, D154 (2008).

[34] R. D. Finn, J. Tate, J. Mistry et al., *Nucleic Acids Research* **36**, D281 (2008).
[35] T. Rattei, P. Tischler, R. Arnold et al., *Nucleic Acids Research* **36**, D289 (2008).
[36] L. Li, C. J. Stoeckert, and D. S. Roos, *Genome Research* **13**, 2178 (2003).
[37] S van Dongen, 2000; A.J. Enright, S van Dongen, and C. A. Ouzounis, *Nucleic Acids Res* **30**, 1575 (2002).
[38] A. H. Paterson, B. A. Chapman, J. Kissinger et al., *Trends Genet.* **22**, 597 (2006).
[39] K. A. Frazer, L. Pachter, A. Poliakov et al., *Nucleic Acids Res* **32**, W273 (2004).
[40] M. Brudno, S. Malde, A. Poliakov et al., *Bioinformatics* **19**, 54i (2003).
[41] S. Schwartz, W. J. Kent, A. Smit et al., *Genome Research* **1**, 103 (2003).
[42] K. Tamura, J. Dudley, M. Nei et al., *Mol Biol Evol* **24** (8), 1596 (2007).
[43] R. Ming, S. Hou, Y. Feng et al., *Nature* **452** (##), 991 (2008).
[44] G. Besnard, B. Offmann, C. Robert et al., *Theor Appl Genet* **105** (2-3), 404 (2002).
[45] R. Wyrich, U. Dressen, S. Brockmann et al., *Plant Mol Biol* **37** (2), 319 (1998).
[46] M. Shenton, V. Fontaine, J. Hartwell et al., *Plant J* **48** (1), 45 (2006).
[47] P. Luchetta, C. Cretin, and P. Gadal, *Mol Gen Genet* **228** (3), 473 (1991).
[48] B. A. Rothermel and T. Nelson, *J Biol Chem* **264** (33), 19587 (1989).
[49] R Chenna, H Sugawara, T. Koike et al., *Nucleic Acids Res* **31**, 3497 (2003); N Saitou and M. Nei, *Mol Biol Evol* **4**, 406 (1987).
[50] F Corpet, *Nucleic Acids Res* **16**, 10881 (1988).
[51] F Chevenet, C. Brun, A.L. Banuls et al., *BMC Bioinformatics* **7**, 439 (2006).