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]

Sorghum, an African grass related to sugar cane and maize, is grown for food, feed, fibre and fuel. We present an initial analysis of the ~730-megabase *Sorghum bicolor* (L.) Moench genome, placing ~98% of genes in their chromosomal context using whole-genome shotgun sequence validated by genetic, physical and syntenic 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 compared with rice. Although gene and repetitive DNA distributions have been preserved since palaeopolyploidization ~70 million years ago, most duplicated gene sets lost one member before the sorghum–rice divergence. Concerted evolution makes one duplicated chromosomal segment appear to be only a few million years old. About 24% of genes are grass-specific and 7% are sorghum-specific. Recent gene and microRNA duplications may contribute to sorghum's drought tolerance.

The Saccharinae plants include some of the most efficient biomass accumulators, providing food and fuel from starch (sorghum) and sugar (sorghum and *Saccharum*, sugar cane), and have potential for use as cellulosic biofuel crops (sorghum, sugar cane, *Miscanthus*). Of singular importance to Saccharinae productivity is C_4 photosynthesis, comprising biochemical and morphological specializations that increase net carbon assimilation at high temperatures[1]. Despite their common photosynthetic strategy, the Saccharinae show much morphological and genomic variation (Supplementary Fig. 1).

Its small genome (~730 Mb) makes sorghum an attractive model for functional genomics of Saccharinae and other C_4 grasses. Rice, the first fully sequenced cereal genome, is more representative of C_3 photosynthetic grasses. Drought tolerance makes sorghum especially important in dry regions such as northeast Africa (its centre of diversity) and the southern plains of the United States. Genetic variation in the partitioning of carbon into sugar stores versus cell wall mass, and in perenniality and associated features such as tillering and stalk reserve retention[2], make sorghum an attractive system for the study of traits important in perennial cellulosic biomass crops. Its high level of inbreeding makes it an attractive association genetics system[3]. Transgenic approaches to sorghum improvement are constrained by high gene flow to weedy relatives[4], making knowledge of its intrinsic genetic potential all the more important.

Reconstructing a repeat-rich genome from shotgun sequences

Preferred approaches to sequencing entire genomes are currently to apply shotgun sequencing[5] either to a minimum 'tiling path' of genomic clones, or to genomic DNA directly. The latter approach, whole-genome shotgun (WGS) sequencing, 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[6].

Despite a repeat content of ~61%, a high-quality genome sequence was assembled from homozygous sorghum genotype BTx623 by using WGS and incorporating the following: (1) ~8.5 genome equivalents of paired-end reads[7] from genomic libraries spanning a ~100-fold range of insert sizes (Supplementary Table 1), resolving many repetitive regions; and (2) high-quality read length averaging 723 bp, facilitating assembly. Comparison with 27 finished bacterial artificial chromosomes (BACs) showed the WGS assembly to be >98.46% complete and accurate to <1 error per 10 kb (Supplementary Note 2.5).

[1]Plant Genome Mapping Laboratory, University of Georgia, Athens, Georgia 30602, USA. [2]Waksman Institute for Microbiology, Rutgers University, Piscataway, New Jersey 08854, USA. [3]DOE Joint Genome Institute, Walnut Creek, California 94598, USA. [4]Stanford Human Genome Center, Stanford University, Palo Alto, California 94304, USA. [5]MIPS/IBIS, Helmholtz Zentrum München, Inglostaedter Landstrasse 1, 85764 Neuherberg, Germany. [6]Center for Integrative Genomics, University of California, Berkeley, California 94720, USA. [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, South Carolina 29631, USA. [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, California 94720, USA. [12]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA. [13]Department of Biological Sciences, [14]Department of Botany and Plant Pathology, Purdue University, West Lafayette, Indiana 47907, USA. [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, Texas 77843, USA. [17]Institute for Genomic Diversity, Cornell University, Ithaca, New York 14853, USA. [18]Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. [19]Mississippi Genome Exploration Laboratory, Mississippi State University, Starkville, Mississippi 39762, USA. [20]National Institute for Biotechnology & Genetic Engineering (NIBGE), Faisalabad, Pakistan. [21]USDA NAA Robert Holley Center for Agriculture and Health, Ithaca, New York 14853, USA.

Comparison with a high-density genetic map[8], a ‘finger-print contig’ (FPC)-based physical map[9], and the rice sequence[6] improved the sorghum WGS assembly (Supplementary Notes 1 and 2). Among the 201 largest scaffolds (spanning 678.9 Mb, 97.3% of the assembly), 28 showed discrepancies with two or more of these lines of evidence (Supplementary Note 2.6), often near repetitive elements. After breaking the assembly at the points of discrepancy, the resulting 229 scaffolds have an N50 (number of scaffolds that collectively cover at least 50% of the assembly) of 35 and L50 (length of the shortest scaffold among those that collectively cover 50% of the assembly) of 7.0 Mb. A total of 38 (2%) of 1,869 FPC contigs[9] were deemed erroneous, containing >5 BAC ends that fell into different sequence scaffolds.

A total of 127 scaffolds containing 625.7 Mb (89.7%) of DNA and 1,476 FPC contigs could be assigned to chromosomal locations and oriented. Fifteen out of twenty chromosome ends terminated in telomeric repeats. The other 102 scaffolds were generally smaller (53.2 Mb, 7.6%), with 85 (83%) containing far greater-than-average abundance of the Cen38 (ref. 10) centromeric repeat, and with only 374 predicted genes. These 102 scaffolds merged only 193 FPC contigs, presumably due to the greater abundance of repeats that are recalcitrant to clone-based physical mapping[9] and may be omitted in BAC-by-BAC approaches[11].

Genome size evolution and its causes

The ~75% larger quantity of DNA in the genome of sorghum compared with rice is mostly heterochromatin. Alignment to genetic[8] and cytological maps[12] suggests that sorghum and rice have similar quantities of euchromatin (252 and 309 Mb, respectively; Supplementary Table 7), accounting for 97–98% of recombination (1,025.2 cM and 1,496.5 cM, respectively) and 75.4–94.2% of genes in the respective cereals, with largely collinear gene order[9]. In contrast, sorghum heterochromatin occupies at least 460 Mb (62%), far more than in rice (63 Mb, 15%). The ~3× genome expansion in maize since its divergence from sorghum[13] has been more dispersed—recombinogenic DNA has grown 4.5× to ~1,382 Mb, much more than can be explained by genome duplication[14].

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

Although recent retroelement activity is widely distributed across the sorghum genome, turnover is rapid (as in other cereals[15]) with pericentromeric elements persisting longer. Young LTR retrotransposon insertions (<0.01 million years (Myr) ago) appear randomly distributed along chromosomes, suggesting that they are preferentially eliminated from gene-rich regions[9] but accumulate in gene-poor regions (Fig. 1; see also Supplementary Note 3.1). Insertion times suggest a major wave of retrotransposition <1 Myr ago, after a smaller wave 1–2 Myr ago (Supplementary Fig. 2).

CACTA-like elements, the predominant sorghum DNA transposons (4.7% of the genome), seem to relocate genes and gene fragments, as do rice ‘Pack-MULEs’[16] and maize helitrons[17]. Many sorghum CACTA elements are non-autonomous deletion derivatives in which transposon genes have been replaced with non-transposon DNA including exons from one or more cellular genes as exemplified for family *G118* (Fig. 2). Among 13,775 CACTA elements identified (Supplementary Note 3.4), 200 encode no transposon proteins but contain at least one cellular gene fragment.

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, 1.7% of the genome, are associated with genes (Fig. 1; see also Supplementary Note 3) as in other cereals[6]. Helitrons, ~0.8% of the genome, nearly all lack helicase in sorghum as in maize[17], but carry fewer gene fragments in sorghum than maize (Supplementary Note 3.5). Organellar DNA insertion has contributed only 0.085% to the sorghum nuclear genome, far less than the 0.53% of rice (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,

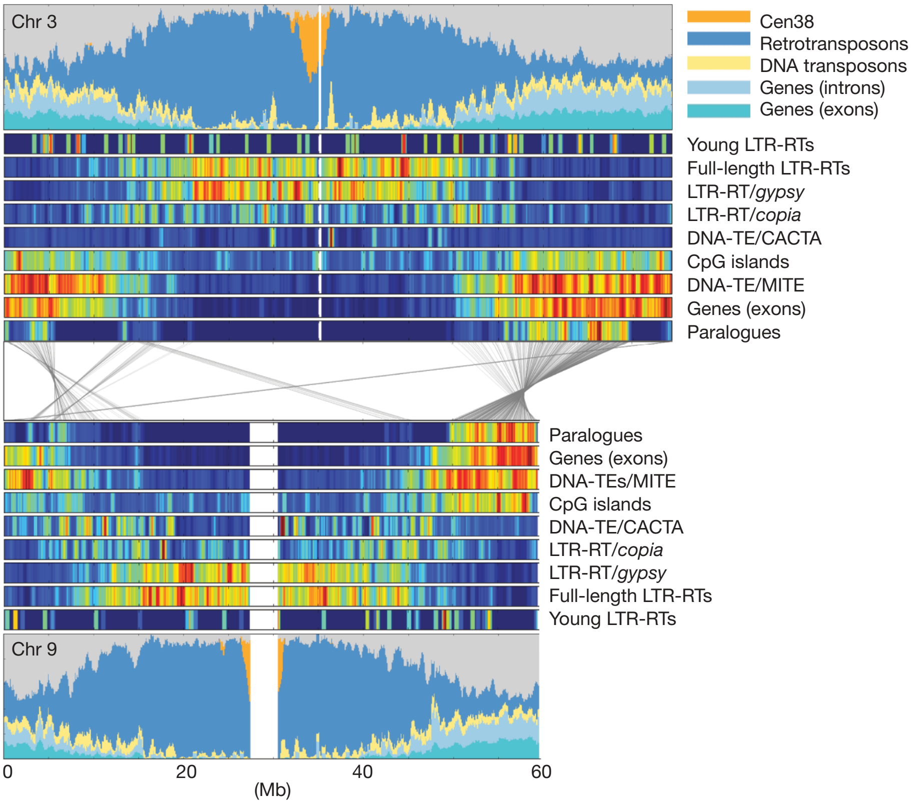

Figure 1 | Genomic landscape of sorghum chromosomes 3 and 9. Area charts quantify retrotransposons (55%), genes (6% exons, 8% introns), DNA transposons (7%) and centromeric repeats (2%). Lines between chromosomes 3 and 9 connect collinear duplicated genes. Heat-map tracks detail the distribution of selected elements. Figures for all sorghum chromosomes are in Supplementary Note 3. Cen38, sorghum-specific centromeric repeat[10]; RTs, retrotransposons (class I); LTR-RTs, long terminal repeat retrotransposons; DNA-TEs, DNA transposons (class II).

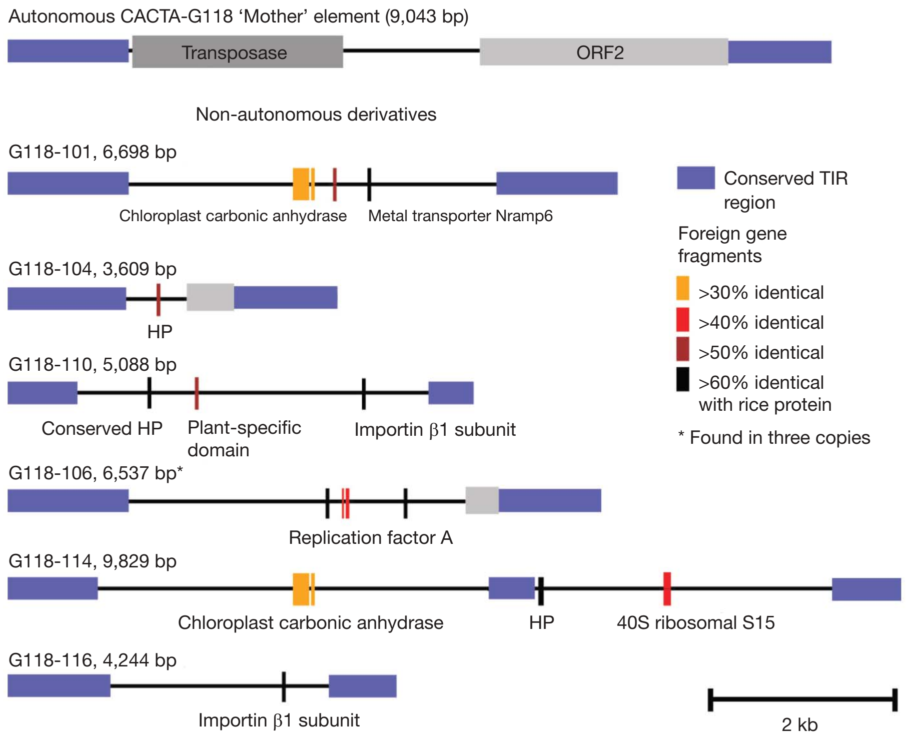

Figure 2 | CACTA element deletion derivatives that carry gene fragments. CACTA family G118 has only one complete and presumably autonomous 'mother' element. Among 18 deletion derivatives, only the terminal 500–2,500 bp are conserved, with 8 carrying gene fragments internally. One relatively homogeneous subgroup (106, 111 and 112) presumably arose recently, whereas other derivatives are unique. 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.

maize and sugar cane (Supplementary Note 4). Evidence for alternate splicing is found in 1,491 loci.

Another 5,197 gene models are usually shorter than the bona fide genes (often <150 amino acids); have few exons (often one) and no expressed sequence tag (EST) support (compared with 85% for bona fide genes); are more diverged from rice genes; and are often found in large families with 'hypothetical', 'uncharacterized' and/or retroelement-associated annotations, despite repeat masking (Supplementary Note 4). A high concentration in pericentromeric regions where bona fide genes are scarce (Fig. 1) suggests that many of these low confidence gene models are retroelement-derived. We also identified 727 processed pseudogenes and 932 models containing domains known only from transposons.

The exon size distributions of orthologous sorghum and rice genes agree closely, and intron position and phase show >98% concordance (Supplementary Note 5). Intron size has been conserved between sorghum and rice, although it has increased in maize owing to transpositions[18].

Most paralogues in sorghum are proximally duplicated, including 5,303 genes in 1,947 families of ≥2 genes (Supplementary Note 4.3). The longest tandem gene array is 15 cytochrome P450 genes. Other sorghum-specific tandem gene expansions include haloacid dehalogenase-like hydrolases (PF00702), FNIP repeats (PF05725), and male sterility proteins (PF03015).

We confirmed the genomic locations of 67 known sorghum microRNAs (miRNAs) and identified 82 additional miRNAs (Supplementary Note 4.4). Five clusters located within 500 bp of each other represent putative polycistronic miRNAs, similar to those in *Arabidopsis* and *Oryza*. Natural antisense miRNA precursors (nat-miRNAs) of family miR444 (ref. 19) have been identified in three copies.

Comparative gene inventories of angiosperms

The number and sizes of sorghum gene families are similar to those of *Arabidopsis*, rice and poplar (Fig. 3 and Supplementary Note 4.6). A total of 9,503 (58%) sorghum gene families were shared among all four species and 15,225 (93%) with at least one other species. Nearly 94% (25,875) of high-confidence sorghum genes have orthologues 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; 1,153 (7%) appear to be unique to sorghum.

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 (Supplementary Table 20). Domains over-represented in sorghum are usually present in the other organisms, a notable exception being the α-kafirin domain that accounts for most seed storage protein and corresponds to maize zeins[20] but which is absent from rice.

Nucleotide-binding-site–leucine-rich-repeat (NBS-LRR) containing proteins associated with the plant immune system are only about half as frequent in sorghum as in rice. A search with 12 NBS domains from published rice, maize, wheat and *Arabidopsis* gene sequences revealed 211 NBS-LRR coding genes in sorghum, 410 in rice and 149 in *Arabidopsis*[21]. Sorghum NBS-LRR genes mostly encode the CC type of N-terminal domains. Only two sorghum genes (Sb02g005860 and Sb02g036630) contain the TIR domain, and neither contains an NBS domain. NBS-LRR genes are most abundant on sorghum chromosome 5 (62), and its rice homologue (chromosome 11, 106). Enrichment of NBS-LRR genes in these corresponding genomic regions suggests conservation of R gene location, in contrast to a proposal that R gene movement may be advantageous[22].

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 orthologues in rice. Additional duplicates formed in maize after the sorghum–maize split (*Zmppck2* and *Zmppck3*). The C_4 NADP-dependent malic enzyme (*me*) gene has an adjacent isoform but each corresponds to a different maize homologue, 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

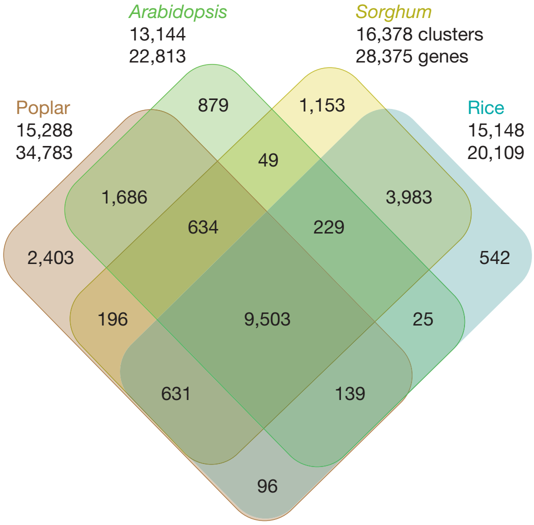

Figure 3 | Orthologous gene families between sorghum, *Arabidopsis*, rice and poplar. The numbers of gene families (clusters) and the total numbers of clustered genes are indicated for each species and species intersection.

copy, suggesting duplication and recruitment to the C_4 pathway after the Panicoideae–Oryzoideae divergence (Supplementary Note 9).

The sorghum sequence reinforces inferences previously based only on rice, about how different grass and dicotyledon gene inventories relate to their respective types of cell walls[23,24]. 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 dicotyledons and rice in the distribution of cell wall biogenesis genes (Supplementary Note 10). For example, the CesA/Csl superfamily and callose synthases have either diverged to form new subgroups or functionally non-essential subgroups were selectively lost, such as *CslB* and *CslG* lost from the grasses, and *CslF* and *CslH* lost from species with dicotyledon-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, respectively.

The characteristic adaptation of sorghum to drought may be partly related to expansion of one miRNA and several gene families. Rice miRNA 169g, upregulated during drought stress[27], has five sorghum homologues (sbi-MIR169c, sbi-MIR169d, 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 by *Arabidopsis* and maize[28]. Cytochrome P450 domain-containing genes, often involved in scavenging toxins such as those accumulated in response to stress, are abundant in sorghum with 326 versus 228 in rice. Expansins, enzymes that break hydrogen bonds and are responsible for a variety of growth responses that could be linked to the durability of sorghum, occur in 82 copies in sorghum versus 58 in rice and 40 each in *Arabidopsis* and poplar.

Duplication and diversification of cereal genomes

Whole-genome duplication in a common ancestor of cereals is reflected in sorghum and rice gene 'quartets' (Fig. 4). A total of 19,929 (57.8%) sorghum gene models were in blocks collinear with rice (Supplementary Note 6). After the shared whole-genome duplication, only one copy was retained for 13,667 (68.6%) collinear genes with 13,526 (99%) being orthologous in rice–sorghum, indicating that most gene losses predate taxon divergence. Both sorghum and rice retained both copies of 4,912 (14.2%) genes, whereas sorghum lost one copy of 1,070 (3.1%) and rice lost one copy of 634 (1.8%). These patterns are likely to be predictive of other grass genomes, as the major grass lineages diverged from a common ancestor at about the same time[29] (see also Supplementary Note 7).

Although 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). Because the sorghum–rice divergence is thought to have happened 20 Myr or more after genome duplication[29], this suggests that even long-term gene loss differentially affects gene functional groups.

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 ~5–7-Myr-old segmental duplication[30–32]. We found a duplicated segment in the corresponding regions of sorghum chromosomes 5 and 8 (Fig. 5). Sorghum–sorghum and rice–rice paralogues from this region show rates of synonymous DNA substitution (K_s) of 0.44 and 0.22, respectively, consistent with only 34 and 17 Myr of divergence. However, the K_s value of sorghum–rice orthologues is 0.63, similar to the respective genome-wide averages (0.81, 0.87). We hypothesize that the apparent segmental duplication actually resulted from the pan-cereal whole-genome duplication and became differentiated from the remainder of the chromosome(s) owing 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 elsewhere in the genome[33]. Physical and genetic maps suggest shared terminal segments of the corresponding chromosomes in wheat (4, 5)[34], foxtail millet (VII, VIII) and pearl millet (linkage groups 1, 4)[35].

Synthesis and implications

Comparison of the sorghum, rice and other genomes clarifies the grass gene set. Pairs of orthologous sorghum and rice genes combined with recent paralogous duplications define 19,542 conserved grass gene families, each representing one gene in the sorghum–rice common ancestor. Our sorghum gene count is similar to that in a manually curated rice annotation (RAP2)[36], but this similarity masks some differences. About 2,054 syntenic orthologues shared by our sorghum annotation and the TIGR5 (ref. 37) rice annotation are absent from RAP2. Conversely, ~12,000 TIGR5 annotations may be transposable elements or pseudogenes, comprising large families of hypothetical genes in both sorghum and rice RAP2, often with short exons, few introns and limited EST support. Phylogenetically incongruent cases of apparent gene loss (for example, genes shared by *Arabidopsis* and sorghum but not rice: Fig. 3) may also suggest sequence gaps or misannotations.

Grass genome architecture may reflect euchromatin-specific effects of recombination and selection, superimposed on non-adaptive processes of mutation and genetic drift that apply to all genomic regions[38]. Patterns of gene and repetitive DNA organization remain correlated in homologous chromosomes duplicated 70 Myr ago (Fig. 1), despite extensive turnover of specific repetitive elements. Synteny is highest and retroelement abundance lowest in distal chromosomal regions. More rapid retroelement removal from gene-rich euchromatin that frequently recombines than from heterochromatin that rarely recombines supports the hypothesis that recombination may preserve gene structure, order and/or spacing by exposing new insertions to selection[9]. Less euchromatin–heterochromatin polarization in maize, where retrotransposon persistence in euchromatin seems more frequent, may reflect variation in grass genome architecture or perhaps a lingering consequence of more recent genome duplication[39].

Identification of conserved DNA sequences may help us to understand essential genes and binding sites that define grasses. Progress in sequencing *Brachypodium distachyon*[40] sets the stage for panicoid–oryzoid–pooid phylogenetic triangulation of genomic changes, as well as association of some such changes with phenotypes ranging from molecular (gene expression patterns) to morphological. The divergence between sorghum, rice and *Brachypodium* is sufficient to randomize

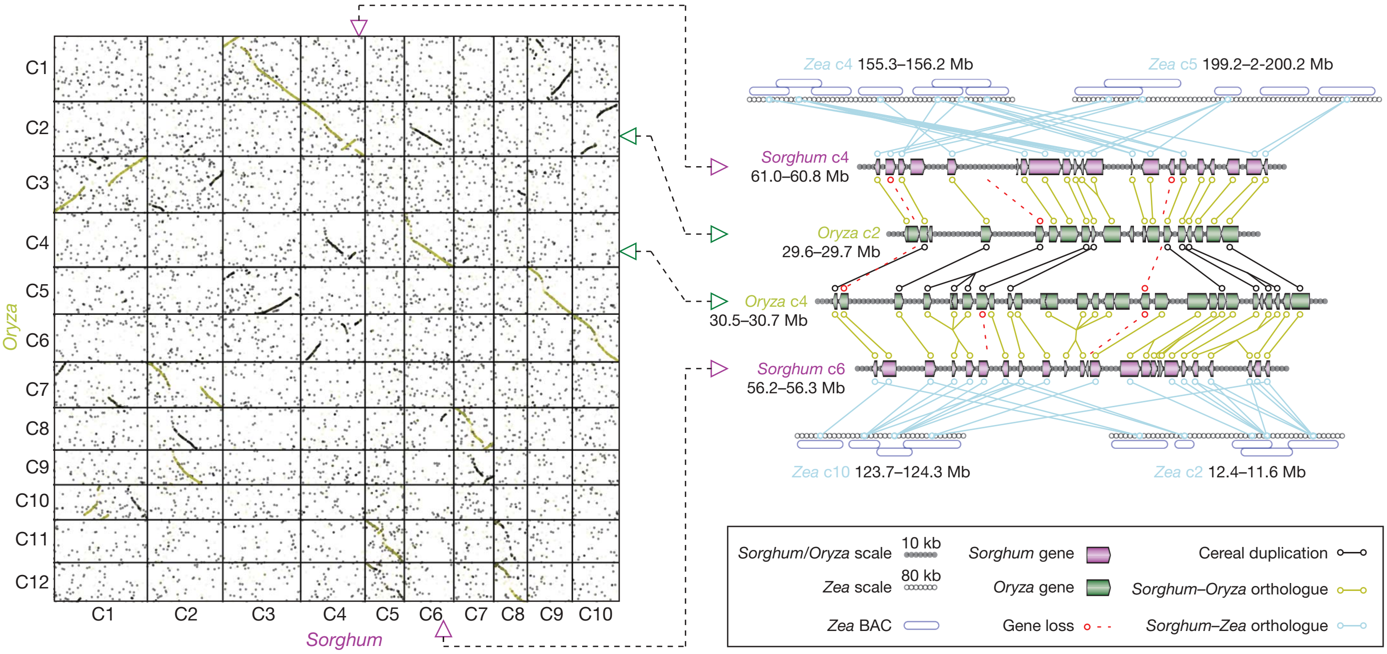

Figure 4 | Alignment of sorghum, rice and maize. Dot plots show intergenomic (gold) and intragenomic (black) alignments. One sorghum–rice quartet showing both orthologous and paralogous (duplicated) regions is magnified. Infrequent gene loss (red; see legend) after sorghum–rice divergence causes 'special cases' in which there are paralogues but no orthologues. Each sorghum region corresponds to two duplicated maize regions[39], with maize gene loss suggested where sorghum loci only match one of the two. Because maize BACs are mostly unfinished, sorghum loci are aligned to the centres. Note the different scale necessary for maize physical distance. Larger dot plots are in Supplementary Note 6.

nonfunctional sequence, facilitating conserved noncoding sequence (CNS) discovery[41,42] (Supplementary Fig. 9). More distant comparisons to the dicotyledon *Arabidopsis* show exon conservation but no CNS (Supplementary Fig. 10). Chloridoid and arundinoid genome sequences are needed to sample the remaining grass lineages, and an outgroup such as *Ananas* (pineapple) or *Musa* (banana) would further aid in identifying genes and sequences that define grasses.

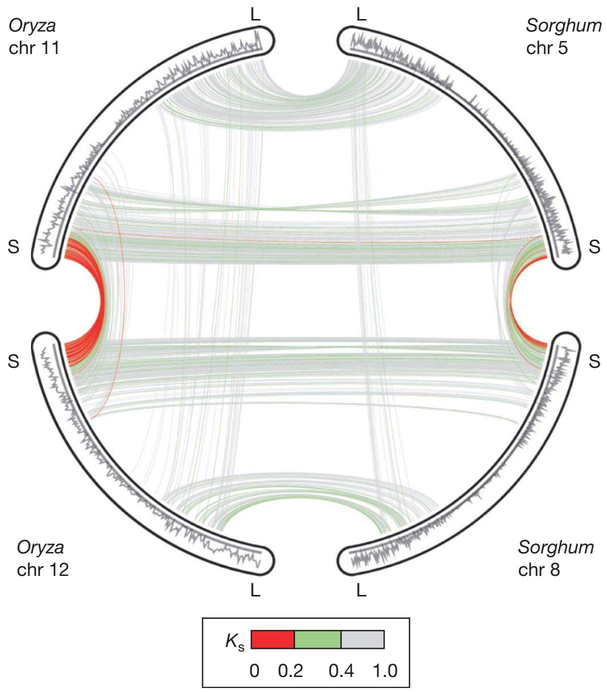

Figure 5 | Independent illegitimate recombination in corresponding regions of sorghum and rice. Four homologous rice and sorghum chromosomes (11 and 12 in rice; 5 and 8 in sorghum) are shown, with gene densities plotted. 'L' and 'S' show long and short arms, respectively. Lines show K_s between homologous gene pairs, and colours are used to show different dates of conversion events.

The fact that the sorghum genome has not re-duplicated in ~70 Myr[29] makes it a valuable outgroup for deducing fates of gene pairs and CNS in grasses that have reduplicated. Single sorghum regions correspond to two regions resulting from maize-specific genome doubling[39]—gene fractionation is evident (Fig. 4), and subfunctionalization is probable (Supplementary Fig. 10). Sorghum may prove especially valuable for unravelling genome evolution in the more closely related *Saccharum–Miscanthus* clade: two genome duplications since its divergence from sorghum 8–9 Myr ago[43] complicate sugar cane genetics[44] yet *Saccharum* BACs show substantially conserved gene order with sorghum (Supplementary Note 11).

Conservation of grass gene structure and order facilitates development of DNA markers to support crop improvement. We identified ~71,000 simple-sequence repeats (SSRs) in sorghum (Supplementary List 1); among a sampling of 212, only 9 (4.2%) map to paralogues of their source locus. Conserved-intron scanning primers (Supplementary List 2) for 6,760 genes provide DNA markers useful across many monocotyledons, particularly valuable for 'orphan cereals'[45].

As the first sequenced plant genome of African origin, sorghum adds new dimensions to ethnobotanical research. Of particular interest will be the identification of alleles selected during the earliest stages of sorghum cultivation, which are valuable towards testing the hypothesis that convergent mutations in corresponding genes contributed to independent domestications of divergent cereals[46]. Invigorated sorghum improvement would benefit regions such as the African 'Sahel' where drought tolerance makes sorghum a staple for human populations that are increasing by 2.8% per year. Sorghum yield improvement has lagged behind that of other grains, in Africa only gaining a total of 37% (western) to 38% (eastern) from 1961–63 to 2005–07 (Supplementary Note 12).

METHODS SUMMARY

Genome sequencing. 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 exceeding 39 kb, excluding 'N's, and collectively representing 678,902,941 bp (97.3%) of nucleotides, were checked for possible chimaeras suggested by the sorghum genetic map, sorghum physical map, abrupt changes in gene or repeat density, rice gene order, and coverage by BAC or fosmid clones (Supplementary Note 2).

Repeat analysis. *De novo* searches for LTR retrotransposons used LTR_STRUC. *De novo* detection of CACTA-DNA transposons and MITEs used custom programs (Supplementary Note 3). Known repeats were identified by RepeatMasker (Open-3-1-8) (http://www.repeatmasker.org) with mips-REdat_6.2_Poaceae, a compilation of grass repeats including sorghum-specific LTR retrotransposons (http://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 alignments of rice and *Arabidopsis* peptides and sorghum and maize ESTs. GenomeScan[47] was applied using maize-specific parameters. Predicted coding structures were merged with EST data from maize and sorghum using PASA[48].

Intergenomic and intragenomic alignments. Dot plots used ColinearScan[49] and multi-alignments used MCScan[50], applied to RAP2[36] (mapped representative models, 29,389 loci) and the sbi1.4 annotation set (34,496 loci). Pairwise BLASTP ($E < 1 \times 10^{-5}$, 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 (http://www.maizesequence.org, release 7 January 2008). *Zea* BACs were searched for potential orthologues of *Sorghum* coding sequences using translated BLAT with a minimum score of 100.

Received 20 August; accepted 9 December 2008.

1. Hatch, M. D. & Slack, C. R. Photosynthesis by sugar-cane leaves—a new carboxylation reaction and pathway of sugar formation. *Biochem. J.* **101,** 103 (1966).
2. Paterson, A. H. *et al.* The weediness of wild plants—molecular analysis of genes influencing dispersal and persistence of johnsongrass, *Sorghum halepense* (l) pers. *Proc. Natl Acad. Sci. USA* **92,** 6127–6131 (1995).
3. Hamblin, M. T. *et al.* Equilibrium processes cannot explain high levels of short- and medium-range linkage disequilibrium in the domesticated grass *Sorghum bicolour*. *Genetics* **171,** 1247–1256 (2005).
4. Morrell, P. L. *et al.* Crop-to-weed introgression has impacted allelic composition of johnsongrass populations with and without recent exposure to cultivated sorghum. *Mol. Ecol.* **14,** 2143–2154 (2005).
5. Gardner, R. C. *et al.* The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. *Nucleic Acids Res.* **9,** 2871–2888 (1981).
6. Matsumoto, T. *et al.* The map-based sequence of the rice genome. *Nature* **436,** 793–800 (2005).
7. Vieira, J. & Messing, J. The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. *Gene* **19,** 259–268 (1982).
8. Bowers, J. E. *et al.* A high-density genetic recombination map of sequence-tagged sites for *Sorghum*, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. *Genetics* **165,** 367–386 (2003).
9. Bowers, J. E. *et al.* Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses. *Proc. Natl Acad. Sci. USA* **102,** 13206–13211 (2005).
10. Miller, J. T. *et al.* Cloning and characterization of a centromere-specific repetitive DNA element from *Sorghum bicolour*. *Theor. Appl. Genet.* **96,** 832–839 (1998).
11. Venter, J. C. *et al.* Shotgun sequencing of the human genome. *Science* **280,** 1540–1542 (1998).
12. Kim, J. S. *et al.* Chromosome identification and nomenclature of *Sorghum bicolour*. *Genetics* **169,** 1169–1173 (2005).
13. Swigonova, Z. *et al.* Close split of sorghum and maize genome progenitors. *Genome Res.* **14,** 1916–1923 (2004).
14. Swigonova, Z. *et al.* On the tetraploid origin of the maize genome. *Comp. Funct. Genomics* **5,** 281–284 (2004).
15. Swigonova, Z., Bennetzen, J. L. & Messing, J. Structure and evolution of the *r/b* chromosomal regions in rice, maize and sorghum. *Genetics* **169,** 891–906 (2005).
16. Jiang, N. *et al.* Pack-mule transposable elements mediate gene evolution in plants. *Nature* **431,** 569–573 (2004).
17. Brunner, S. *et al.* Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell* **17,** 343–360 (2005).
18. Haberer, G. *et al.* Structure and architecture of the maize genome. *Plant Physiol.* **139,** 1612–1624 (2005).
19. Lu, C. *et al.* Genome-wide analysis for discovery of rice microRNAs reveals natural antisense microRNAs (nat-miRNAs). *Proc. Natl Acad. Sci. USA* **105,** 4951–4956 (2008).
20. Xu, J.-H. & Messing, J. Organization of the prolamin gene family provides insight into the evolution of the maize genome and gene duplications in grass species. *Proc. Natl Acad. Sci. USA* **105,** 14330–14335 (2008).
21. Meyers, B. C. *et al.* Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell* **15,** 809–834 (2003).
22. Leister, D. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. *Trends Genet.* **20,** 116–122 (2004).
23. Carpita, N. C. & Gibeaut, D. M. Structural models of primary cell walls in flowering plants—consistency of molecular structure with the physical properties of the walls during growth *Plant J.* **3,** 1–30 (1993).
24. McCann, M. C. & Roberts, K. in *The Cytoskeletal Basis of Plant Growth and Form* (ed. Lloyd, C. W.) 109–129 (Academic Press, 1991).
25. Carpita, N. C. Structure and biogenesis of the cell walls of grasses. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **47,** 445–476 (1996).
26. Hazen, S. P. *et al.* Quantitative trait loci and comparative genomics of cereal cell wall composition. *Plant Physiol.* **132,** 263–271 (2003).
27. Zhao, B. T. *et al.* Identification of drought-induced microRNAs in rice. *Biochem. Biophys. Res. Commun.* **354,** 585–590 (2007).
28. Nelson, D. E. *et al.* Plant nuclear factor Y (NF-Y) B subunits confer drought tolerance and lead to improved corn yields on water-limited acres *Proc. Natl Acad. Sci. USA* **104,** 16450–16455 (2007).
29. Paterson, A. H., Bowers, J. E. & Chapman, B. A. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl Acad. Sci. USA* **101,** 9903–9908 (2004).
30. Wang, X. *et al.* Duplication and DNA segmental loss in rice genome and their implications for diploidization. *New Phytol.* **165,** 937–946 (2005).
31. Yu, J. *et al.* The genomes of *Oryza sativa*: A history of duplications. *PLoS Biol.* **3,** 266–281 (2005).
32. The Rice Chromosomes 11 and 12 Sequencing Consortia.. The sequence of rice chromosomes 11 and 12, rich in disease resistance genes and recent gene duplications. *BMC Biol.* **3,** 20 (2005).
33. Wang, X. *et al.* Extensive concerted evolution of rice paralogs and the road to regaining independence. *Genetics* **177,** 1753–1763 (2007).
34. Singh, N. K. *et al.* Single-copy genes define a conserved order between rice and wheat for understanding differences caused by duplication, deletion, and transposition of genes. *Funct. Integr. Genomics* **7,** 17–35 (2007).
35. Devos, K. M., Pittaway, T. S., Reynolds, A. & Gale, M. D. Comparative mapping reveals a complex relationship between the pearl millet genome and those of foxtail millet and rice *TAG. Theor. Appl. Genet.* **100,** 190–198 (2000).
36. Tanaka, T. *et al.* The rice annotation project database (RAP-DB): 2008 update. *Nucleic Acids Res.* **36,** D1028–D1033 (2008).
37. Ouyang, S. *et al.* The TIGR rice genome annotation resource: Improvements and new features. *Nucleic Acids Res.* **35,** D883–D887 (2007).
38. Lynch, M. & Conery, J. S. The origins of genome complexity. *Science* **302,** 1401–1404 (2003).
39. Wei, F. *et al.* Physical and genetic structure of the maize genome reflects its complex evolutionary history. *PLoS Genet.* **3,** e123 (2007).
40. Huo, N. *et al.* The nuclear genome of *Brachypodium distachyon*: Analysis of BAC end sequences. *Funct. Integr. Genomics* **8,** 135–147 (2007).
41. Margulies, E. H. *et al.* An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl Acad. Sci. USA* **102,** 4795–4800 (2005).
42. Eddy, S. R. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.* **3,** 95–102 (2005).
43. Jannoo, N. *et al.* Orthologous comparison in a gene-rich region among grasses reveals stability in the sugarcane polyploid genome. *Plant J.* **50,** 574–585 (2007).
44. Ming, R. *et al.* Sugarcane improvement through breeding and biotechnology. *Plant Breed. Rev.* **27,** 15–118 (2005).
45. Lohithaswa, H. C. *et al.* Leveraging the rice genome sequence for comparative genomics in monocots. *Theor. Appl. Genet.* **115,** 237–243 (2007).
46. Paterson, A. H. *et al.* Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. *Science* **269,** 1714–1718 (1995).
47. Yeh, R.-F., Lim, L. P. & Burge, C. Computational inference of homologous gene structures in the human genome. *Genome Res.* **11,** 803–816 (2001).
48. Haas, B. J. *et al.* Full-length messenger RNA sequences greatly improve genome annotation. *Genome Biol.* **3,** research0029.0021–0029.0012 (2002).
49. Wang, X. Y. *et al.* Statistical inference of chromosomal homology based on gene colinearity and applications to *Arabidopsis* and rice. *BMC Bioinform.* **7,** 447 (2006).
50. Tang, H. *et al.* Synteny and colinearity in plant genomes. *Science* **320,** 486–488 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the US Department of Energy Joint Genome Institute Community Sequencing Program, J. Bristow, S. Lucas and the JGI production sequencing team for sequencing sorghum; and L. Lin for contributions to Fig. 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 A.H.P.; US Department of Energy (DE-FG05-95ER20194) to J.M.; German Federal Ministry of Education GABI initiative to MIPS (0313117 and 0314000C); NSF DBI-0321467 to A.N.; and US Department of Agriculture-Agricultural Research Service to C.A.M., L.Z. and D.W.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.H.P. (paterson@uga.edu).