2-2010

Genome sequencing and analysis of the model grass *Brachypodium distachyon*

John Vogel
USDA-ARS, john.vogel@ars.usda.gov

David Garvin
USDA-ARS, david.garvin@ars.usda.gov

Todd C. Mockler
Oregon State University, tmockler@danforthcenter.org

Jeremy Schmutz
International Brachypodium Initiative

Dan Rokhsar
The International Brachypodium Initiative

See next page for additional authors

Vol 463 | 11 February 2010 | doi:10.1038/nature08747 nature

Genome sequencing and analysis of the model grass *Brachypodium distachyon*

The International Brachypodium Initiative*

Three subfamilies of grasses, the Ehrhartoideae, Panicoideae and Pooideae, provide the bulk of human nutrition and are poised to become major sources of renewable energy. Here we describe the genome sequence of the wild grass *Brachypodium distachyon* (*Brachypodium*), which is, to our knowledge, the first member of the Pooideae subfamily to be sequenced. Comparison of the *Brachypodium*, rice and sorghum genomes shows a precise history of genome evolution across a broad diversity of the grasses, and establishes a template for analysis of the large genomes of economically important pooid grasses such as wheat. The high-quality genome sequence, coupled with ease of cultivation and transformation, small size and rapid life cycle, will help *Brachypodium* reach its potential as an important model system for developing new energy and food crops.

Grasses provide the bulk of human nutrition, and highly productive grasses are promising sources of sustainable energy[1]. The grass family (Poaceae) comprises over 600 genera and more than 10,000 species that dominate many ecological and agricultural systems[2,3]. So far, genomic efforts have largely focused on two economically important grass subfamilies, the Ehrhartoideae (rice) and the Panicoideae (maize, sorghum, sugarcane and millets). The rice[4] and sorghum[5] genome sequences and a detailed physical map of maize[6] showed extensive conservation of gene order[5,7] and both ancient and relatively recent polyploidization.

Most cool season cereal, forage and turf grasses belong to the Pooideae subfamily, which is also the largest grass subfamily. The genomes of many pooids are characterized by daunting size and complexity. For example, the bread wheat genome is approximately 17,000 megabases (Mb) and contains three independent genomes[8]. This has prohibited genome-scale comparisons spanning the three most economically important grass subfamilies.

Brachypodium, a member of the Pooideae subfamily, is a wild annual grass endemic to the Mediterranean and Middle East[9] that has promise as a model system. This has led to the development of highly efficient transformation[10,11], germplasm collections[12–14], genetic markers[14], a genetic linkage map[15], bacterial artificial chromosome (BAC) libraries[16,17], physical maps[18] (M.F., unpublished observations), mutant collections (http://brachypodium.pw.usda.gov, http://www.brachytag.org), microarrays and databases (http://www.brachybase.org, http://www.phytozome.net, http://www.modelcrop.org, http://mips.helmholtz-muenchen.de/plant/index.jsp) that are facilitating the use of *Brachypodium* by the research community. The genome sequence described here will allow *Brachypodium* to act as a powerful functional genomics resource for the grasses. It is also an important advance in grass structural genomics, permitting, for the first time, whole-genome comparisons between members of the three most economically important grass subfamilies.

Genome sequence assembly and annotation

The diploid inbred line Bd21 (ref. 19) was sequenced using whole-genome shotgun sequencing (Supplementary Table 1). The ten largest scaffolds contained 99.6% of all sequenced nucleotides (Supplementary Table 2). Comparison of these ten scaffolds with a genetic map (Supplementary Fig. 1) detected two false joins and created a further seven joins to produce five pseudomolecules that spanned 272 Mb (Supplementary Table 3), within the range measured by flow cytometry[20,21]. The assembly was confirmed by cytogenetic analysis (Supplementary Fig. 2) and alignment with two physical maps and sequenced BACs (Supplementary Data). More than 98% of expressed sequence tags (ESTs) mapped to the sequence assembly, consistent with a near-complete genome (Supplementary Table 4 and Supplementary Fig. 3). Compared to other grasses, the *Brachypodium* genome is very compact, with retrotransposons concentrated at the centromeres and syntenic breakpoints (Fig. 1). DNA transposons and derivatives are broadly distributed and primarily associated with gene-rich regions.

We analysed small RNA populations from inflorescence tissues with deep Illumina sequencing, and mapped them onto the genome sequence (Fig. 2a, Supplementary Fig. 4 and Supplementary Table 5). Small RNA reads were most dense in regions of high repeat density, similar to the distribution reported in *Arabidopsis*[22]. We identified 413 and 198 21- and 24-nucleotide phased short interfering RNA (siRNA) loci, respectively. Using the same algorithm, the only phased loci identified in *Arabidopsis* were five of the eight *trans*-acting siRNA loci, and none was 24-nucelotide phased. The biological functions of these clusters of *Brachypodium* phased siRNAs, which account for a significant number of small RNAs that map outside repeat regions, are not known at present.

A total of 25,532 protein-coding gene loci was predicted in the v1.0 annotation (Supplementary Information and Supplementary Table 6). This is in the same range as rice (RAP2, 28,236)[23] and sorghum (v1.4, 27,640)[5], suggesting similar gene numbers across a broad diversity of grasses. Gene models were evaluated using ~10.2 gigabases (Gb) of Illumina RNA-seq data (Supplementary Fig. 5)[24]. Overall, 92.7% of predicted coding sequences (CDS) were supported by Illumina data (Fig. 2b), demonstrating the high accuracy of the *Brachypodium* gene predictions. These gene models are available from several databases (such as http://www.brachybase.org, http://www.phytozome.net, http://www.modelcrop.org and http://mips.org).

Between 77 and 84% of gene families (defined according to Supplementary Fig. 6) are shared among the three grass subfamilies represented by *Brachypodium*, rice and sorghum, reflecting a relatively

*A list of participants and their affiliations appears at the end of the paper.

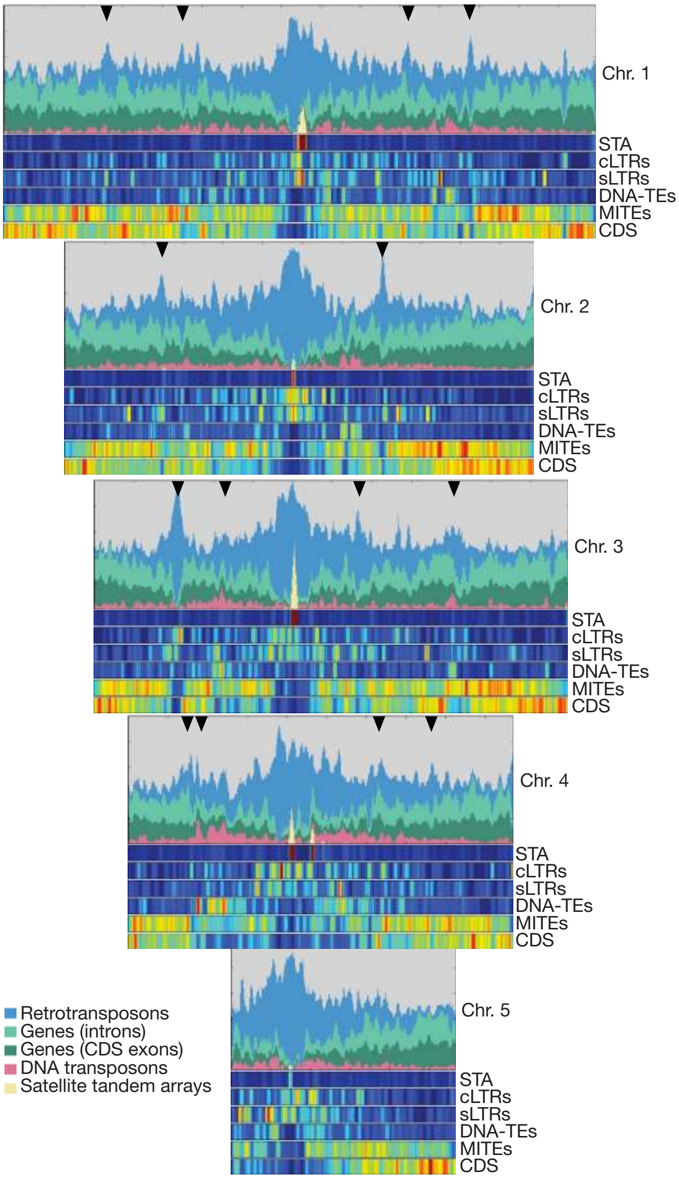


Figure 1 | Chromosomal distribution of the main *Brachypodium* genome features. The abundance and distribution of the following genome elements are shown: complete LTR retroelements (cLTRs); solo-LTRs (sLTRs); potentially autonomous DNA transposons not miniature inverted-repeat transposable elements (MITEs) (DNA-TEs); MITEs; gene exons (CDS); gene introns and satellite tandem arrays (STA). Graphs are from 0 to 100 per cent base-pair (%bp) coverage of the respective window. The heat map tracks have different ranges and different maximum (max) pseudocolour levels: STA (0–55, scaled to max 10) %bp; cLTRs (0–36, scaled to max 20) %bp; sLTRs (0–4) %bp; DNA-TEs (0–20) %bp; MITEs (0–22) %bp; CDS (exons) (0–22.3) %bp. The triangles identify syntenic breakpoints.

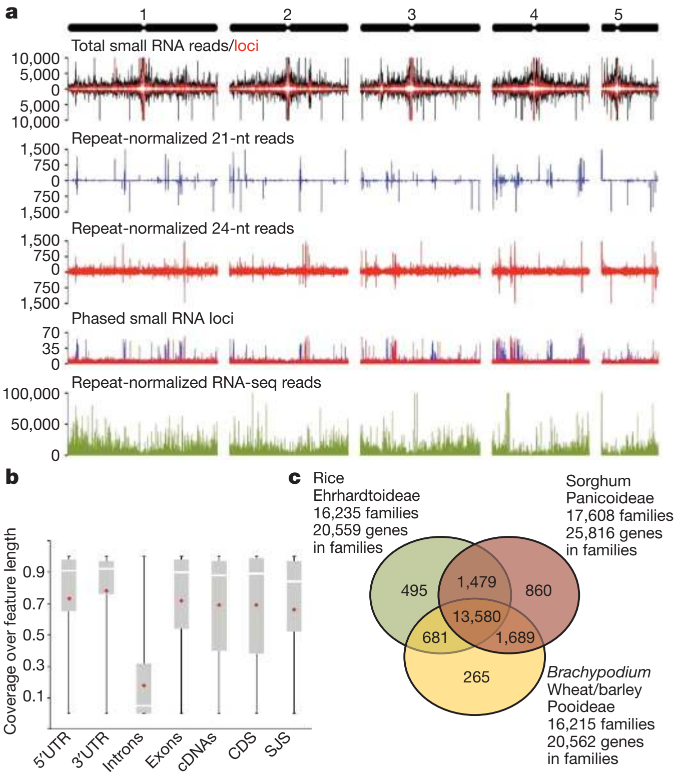


Figure 2 | Transcript and gene identification and distribution among three grass subfamilies. a, Genome-wide distribution of small RNA loci and transcripts in the *Brachypodium* genome. *Brachypodium* chromosomes (1–5) are shown at the top. Total small RNA reads (black lines) and total small RNA loci (red lines) are shown on the top panel. Histograms plot 21-nucleotide (nt) (blue) or 24-nucleotide (red) small RNA reads normalized for repeated matches to the genome. The phased loci histograms plot the position and phase-score of 21-nucleotide (blue) and 24-nucleotide (red) phased small RNA loci. Repeat-normalized RNA-seq read histograms plot the abundance of reads matching RNA transcripts (green), normalized for ambiguous matches to the genome. **b,** Transcript coverage over gene features. Perfect match 32-base oligonucleotide Illumina reads were mapped to the *Brachypodium* v1.0 annotation features using HashMatch (http://mocklerlab-tools.cgrb.oregonstate.edu/). Plots of Illumina coverage were calculated as the percentage of bases along the length of the sequence feature supported by Illumina reads for the indicated gene model features. The bottom and top of the box represent the 25th and 75th quartiles, respectively. The white line is the median and the red diamonds denote the mean. SJS, splice junction site. **c,** Venn diagram showing the distribution of shared gene families between representatives of Ehrhartoideae (rice RAP2), Panicoideae (sorghum v1.4) and Pooideae (*Brachypodium* v1.0, and *Triticum aestivum* and *Hordeum vulgare* TCs (transcript consensus)/EST sequences). Paralogous gene families were collapsed in these data sets.

recent common origin (Fig. 2c). Grass-specific genes include transmembrane receptor protein kinases, glycosyltransferases, peroxidases and P450 proteins (Supplementary Table 7B). The Pooideae-specific gene set contains only 265 gene families (Supplementary Table 7C) comprising 811 genes (1,400 including singletons). Genes enriched in grasses were significantly more likely to be contained in tandem arrays than random genes, demonstrating a prominent role for tandem gene expansion in the evolution of grass-specific genes (Supplementary Fig. 7 and Supplementary Table 8).

To validate and improve the v1.0 gene models, we manually annotated 2,755 gene models from 97 diverse gene families (Supplementary Tables 9–11) relevant to bioenergy and food crop improvement. We annotated 866 genes involved in cell wall biosynthesis/modification and 948 transcription factors from 16 families[25]. Only 13% of the gene models required modification and very few pseudogenes were identified, demonstrating the accuracy of the v1.0 annotation. Phylogenetic trees for 62 gene families were constructed using genes from rice, *Arabidopsis*, sorghum and poplar. In nearly all cases, *Brachypodium* genes had a similar distribution to rice and sorghum, demonstrating that *Brachypodium* is suitably generic for grass functional genomics research (Supplementary Figs 8 and 9). Analysis of the predicted secretome identified substantial differences in the distribution of cell wall metabolism genes between dicots and grasses (Supplementary Tables 12, 13 and Supplementary Fig. 10), consistent with their different cell walls[26]. Signal peptide probability curves also suggested that start codons were accurately predicted (Supplementary Fig. 11).

Maintaining a small grass genome size

Exhaustive analysis of transposable elements (Supplementary Information and Supplementary Table 14) showed retrotransposon sequences comprise 21.4% of the genome, compared to 26% in rice, 54% in sorghum, and more than 80% in wheat[27]. Thirteen retroelement sets were younger than 20,000 years, showing a recent activation compared to rice[28] (Supplementary Fig. 12), and a further 53 retroelement sets were less than 0.1 million years (Myr) old. A minimum of 17.4 Mb has been lost by long terminal repeat (LTR)–LTR recombination, demonstrating that retroelement expansion is countered by removal through recombination. In contrast, retroelements persist for very long periods of time in the closely related Triticeae[28].

DNA transposons comprise 4.77% of the *Brachypodium* genome, within the range found in other grass genomes[5,29]. Transcriptome data and structural analysis suggest that many non-autonomous *Mariner DTT* and *Harbinger* elements recruit transposases from other families. Two *CACTA DTC* families (M and N) carried five non-element genes, and the *Harbinger U* family has amplified a NBS-LRR gene family (Supplementary Figs 13 and 14), adding it to the group of transposable elements implicated in gene mobility[30,31]. Centromeric regions were characterized by low gene density, characteristic repeats and retroelement clusters (Supplementary Fig. 15). Other repeat classes are described in Supplementary Table 15. Conserved non-coding sequences are described in Supplementary Fig. 16.

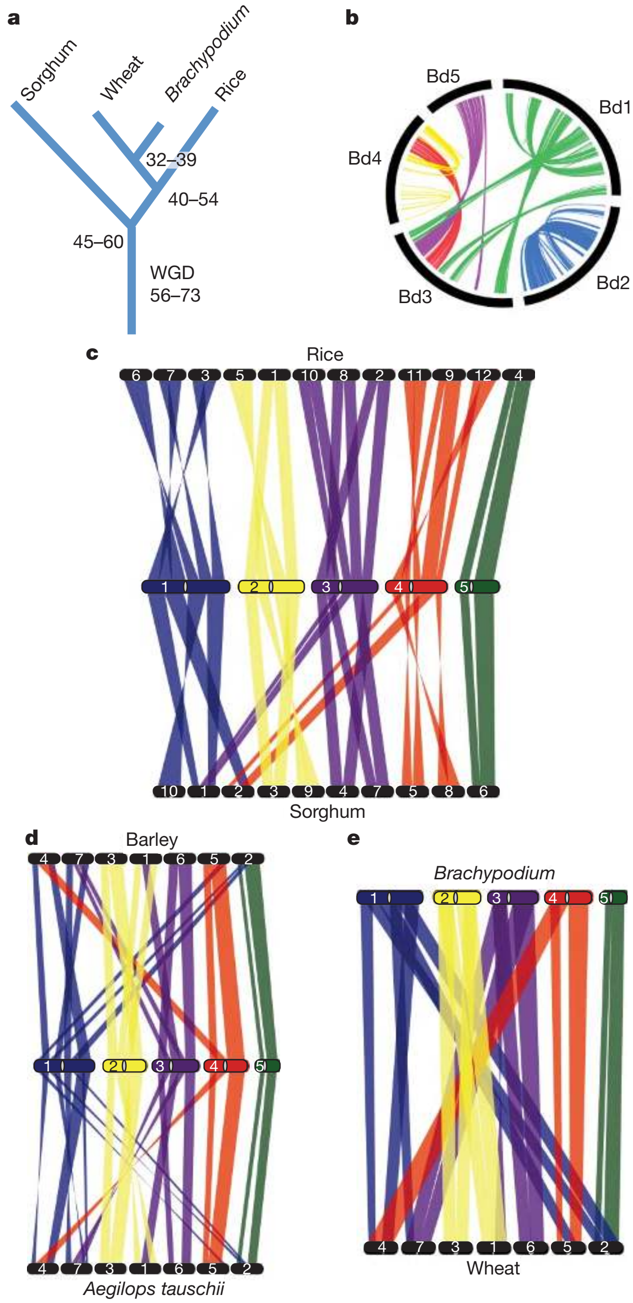


Figure 3 | *Brachypodium* genome evolution and synteny between grass subfamilies. a, The distribution maxima of mean synonymous substitution rates (K_s) of *Brachypodium*, rice, sorghum and wheat orthologous gene pairs (Supplementary Table 16) were used to define the divergence times of these species and the age of interchromosomal duplications in *Brachypodium*. WGD, whole-genome duplication. The numbers refer to the predicted divergence times measured as Myr ago by the NG or ML methods. **b,** Diagram showing the six major interchromosomal *Brachypodium* duplications, defined by 723 paralogous relationships, as coloured bands linking the five chromosomes. **c,** Identification of chromosome relationships between the *Brachypodium*, rice and sorghum genomes. Orthologous relationships between the 25,532 protein-coding *Brachypodium* genes, 7,216 sorghum orthologues (12 syntenic blocks), and 8,533 rice orthologues (12 syntenic blocks) were defined. Sets of collinear orthologous relationships are represented by a coloured band according to each *Brachypodium* chromosome (blue, chromosome (chr.) 1; yellow, chr. 2; violet, chr. 3; red, chr. 4; green, chr. 5). The white region in each *Brachypodium* chromosome represents the centromeric region. **d,** Orthologous gene relationships between *Brachypodium* and barley and *Ae. tauschii* were identified using genetically mapped ESTs. 2,516 orthologous relationships defined 12 syntenic blocks. These are shown as coloured bands. **e,** Orthologous gene relationships between *Brachypodium* and hexaploid bread wheat defined by 5,003 ESTs mapped to wheat deletion bins. Each set of orthologous relationships is represented by a band that is evenly spread across each deletion interval on the wheat chromosomes.

Whole-genome comparison of three diverse grass genomes

The evolutionary relationships between *Brachypodium*, sorghum, rice and wheat were assessed by measuring the mean synonymous substitution rates (K_s) of orthologous gene pairs (Supplementary Information, Supplementary Fig. 17 and Supplementary Table 16), from which divergence times of *Brachypodium* from wheat 32–39 Myr ago, rice 40–53 Myr ago, and sorghum 45–60 Myr ago (Fig. 3a) were estimated. The K_s of orthologous gene pairs in the intragenomic *Brachypodium* duplications (Fig. 3b) suggests duplication 56–72 Myr ago, before the diversification of the grasses. This is consistent with previous evolutionary histories inferred from a small number of genes[3,32–34].

Paralogous relationships among *Brachypodium* chromosomes showed six major chromosomal duplications covering 92.1% of the genome (Fig. 3b), representing ancestral whole-genome duplication[35]. Using the rice and sorghum genome sequences, genetic maps of barley[36] and *Aegilops tauschii* (the D genome donor of hexaploid wheat)[37], and bin-mapped wheat ESTs[38,39], 21,045 orthologous relationships between *Brachypodium*, rice, sorghum and Triticeae were identified (Supplementary Information). These identified 59 blocks of collinear genes covering 99.2% of the *Brachypodium* genome (Fig. 3c–e). The orthologous relationships are consistent with an evolutionary model that shaped five *Brachypodium* chromosomes from a five-chromosome ancestral genome by a 12-chromosome intermediate involving seven major chromosome fusions[39] (Supplementary Fig. 18). These collinear blocks of orthologous genes provide a robust and precise sequence framework for understanding grass genome evolution and aiding the assembly of sequences from other pooid grasses. We identified 14 major syntenic disruptions between *Brachypodium* and rice/sorghum that can be explained by nested insertions of entire chromosomes into centromeric regions (Fig. 4a, b)[2,37,40]. Similar nested insertions in sorghum[37] and barley (Fig. 4c, d) were also identified. Centromeric repeats and peaks in retroelements at the junctions of chromosome insertions are footprints of these insertion events (Supplementary Fig. 15C and Fig. 1), as is higher gene density at the former distal regions of the inserted chromosomes (Fig. 1). Notably, the reduction in chromosome number in *Brachypodium* and wheat occurred independently because none of the chromosome fusions are shared by *Brachypodium* and the Triticeae[37] (Supplementary Fig. 18).

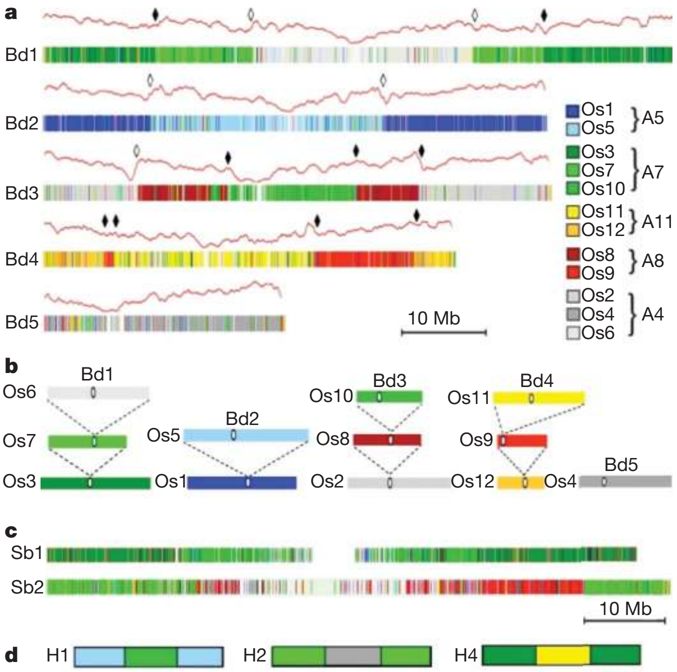


Figure 4 | A recurring pattern of nested chromosome fusions in grasses. **a**, The five *Brachypodium* chromosomes are coloured according to homology with rice chromosomes (Os1–Os12). Chromosomes descended from an ancestral chromosome (A4–A11) through whole-genome duplication are shown in shades of the same colour. Gene density is indicated as a red line above the chromosome maps. Major discontinuities in gene density identify syntenic breakpoints, which are marked by a diamond. White diamonds identify fusion points containing remnant centromeric repeats. **b**, A pattern of nested insertions of whole chromosomes into centromeric regions explains the observed syntenic break points. Bd5 has not undergone chromosome fusion. **c**, Examples of nested chromosome insertions in sorghum (Sb) chromosomes 1 and 2. **d**, Examples of nested chromosome insertions in barley (H chromosomes) inferred from genetic maps. Nested insertions were not identified in other chromosomes, possibly owing to the low resolution of genetic maps.

Comparisons of evolutionary rates between *Brachypodium*, sorghum, rice and *Ae. tauschii* demonstrated a substantially higher rate of genome change in *Ae. tauschii* (Supplementary Table 17). This may be due to retroelement activity that increases syntenic disruptions, as proposed for chromosome 5S later[41]. Among seven relatively large gene families, four were highly syntenic and two (NBS-LRR and F-box) were almost never found in syntenic order when compared to rice and sorghum (Supplementary Table 18), consistent with the rapid diversification of the NBS-LRR and F-box gene families[42].

The short arm of chromosome 5 (Bd5S) has a gene density roughly half of the rest of the genome, high LTR retrotransposon density, the youngest intact *Gypsy* elements and the lowest solo LTR density. Thus, unlike the rest of the *Brachypodium* genome, Bd5S is gaining retrotransposons by replication and losing fewer by recombination. Syntenic regions of rice (Os4S) and sorghum (Sb6S) demonstrate maintenance of this high repeat content for ~50–70 Myr (Supplementary Fig. 19)[43]. Bd5S, Os4S and Sb6S also have the lowest proportion of collinear genes (Fig. 4a and Supplementary Fig. 19). We propose that the chromosome ancestral to Bd5S reached a tipping point in which high retrotransposon density had deleterious effects on genes.

Discussion

As the first genome sequence of a pooid grass, the *Brachypodium* genome aids genome analysis and gene identification in the large and complex genomes of wheat and barley, two other pooid grasses that are among the world's most important crops. The very high quality of the *Brachypodium* genome sequence, in combination with those from two other grass subfamilies, enabled reconstruction of chromosome evolution across a broad diversity of grasses. This analysis contributes to our understanding of grass diversification by explaining how the varying chromosome numbers found in the major grass subfamilies derive from an ancestral set of five chromosomes by nested insertions of whole chromosomes into centromeres. The relatively small genome of *Brachypodium* contains many active retroelement families, but recombination between these keeps genome expansion in check. The short arm of chromosome 5 deviates from the rest of the genome by exhibiting a trend towards genome expansion through increased retroelement numbers and disruption of gene order more typical of the larger genomes of closely related grasses.

Grass crop improvement for sustainable fuel[44] and food[45] production requires a substantial increase in research in species such as *Miscanthus*, switchgrass, wheat and cool season forage grasses. These considerations have led to the rapid adoption of *Brachypodium* as an experimental system for grass research. The similarities in gene content and gene family structure between *Brachypodium*, rice and sorghum support the value of *Brachypodium* as a functional genomics model for all grasses. The *Brachypodium* genome sequence analysis reported here is therefore an important advance towards securing sustainable supplies of food, feed and fuel from new generations of grass crops.

METHODS SUMMARY

Genome sequencing and assembly. Sanger sequencing was used to generate paired-end reads from 3 kb, 8 kb, fosmid (35 kb) and BAC (100 kb) clones to generate 9.4× coverage (Supplementary Table 1). The final assembly of 83 scaffolds covers 271.9 Mb (Supplementary Table 3). Sequence scaffolds were aligned to a genetic map to create pseudomolecules covering each chromosome (Supplementary Figs 1 and 2).

Protein-coding gene annotation. Gene models were derived from weighted consensus prediction from several *ab initio* gene finders, optimal spliced alignments of ESTs and transcript assemblies, and protein homology. Illumina transcriptome sequence was aligned to predicted genome features to validate exons, splice sites and alternatively spliced transcripts.

Repeats analysis. The MIPS ANGELA pipeline was used to integrate analyses from expert groups. LTR-STRUCT and LTR-HARVEST[46] were used for *de novo* retroelement searches.

Received 29 August; accepted 9 December 2009.

1. Somerville, C. The billion-ton biofuels vision. *Science* **312,** 1277 (2006).
2. Kellogg, E. A. Evolutionary history of the grasses. *Plant Physiol.* **125,** 1198–1205 (2001).
3. Gaut, B. S. Evolutionary dynamics of grass genomes. *New Phytol.* **154,** 15–28 (2002).
4. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436,** 793–800 (2005).
5. Paterson, A. H. *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457,** 551–556 (2009).
6. Wei, F. *et al.* Physical and genetic structure of the maize genome reflects its complex evolutionary history. *PLoS Genet.* **3,** e123 (2007).
7. Moore, G., Devos, K. M., Wang, Z. & Gale, M. D. Cereal genome evolution. Grasses, line up and form a circle. *Curr. Biol.* **5,** 737–739 (1995).
8. Salamini, F., Ozkan, H., Brandolini, A., Schafer-Pregl, R. & Martin, W. Genetics and geography of wild cereal domestication in the near east. *Nature Rev. Genet.* **3,** 429–441 (2002).
9. Draper, J. *et al. Brachypodium distachyon*. A new model system for functional genomics in grasses. *Plant Physiol.* **127,** 1539–1555 (2001).
10. Vain, P. *et al.* Agrobacterium-mediated transformation of the temperate grass *Brachypodium distachyon* (genotype Bd21) for T-DNA insertional mutagenesis. *Plant Biotechnol. J.* **6,** 236–245 (2008).
11. Vogel, J. & Hill, T. High-efficiency *Agrobacterium*-mediated transformation of *Brachypodium distachyon* inbred line Bd21-3. *Plant Cell Rep.* **27,** 471–478 (2008).
12. Vogel, J. P., Garvin, D. F., Leong, O. M. & Hayden, D. M. *Agrobacterium*-mediated transformation and inbred line development in the model grass *Brachypodium distachyon*. *Plant Cell Tissue Organ Cult.* **84,** 100179–100191 (2006).
13. Filiz, E. *et al.* Molecular, morphological and cytological analysis of diverse *Brachypodium distachyon* inbred lines. *Genome* **52,** 876–890 (2009).
14. Vogel, J. P. *et al.* Development of SSR markers and analysis of diversity in Turkish populations of *Brachypodium distachyon*. *BMC Plant Biol.* **9,** 88 (2009).

15. Garvin, D. F. *et al.* An SSR-based genetic linkage map of the model grass *Brachypodium distachyon*. *Genome* **53**, 1–13 (2009).

16. Huo, N. *et al.* Construction and characterization of two BAC libraries from *Brachypodium distachyon*, a new model for grass genomics. *Genome* **49**, 1099–1108 (2006).

17. Huo, N. *et al.* The nuclear genome of *Brachypodium distachyon*: analysis of BAC end sequences. *Funct. Integr. Genomics* **8**, 135–147 (2008).

18. Gu, Y. Q. *et al.* A BAC-based physical map of *Brachypodium distachyon* and its comparative analysis with rice and wheat. *BMC Genomics* **10**, 496 (2009).

19. Garvin, D. F. *et al.* Development of genetic and genomic research resources for *Brachypodium distachyon*, a new model system for grass crop research. *Crop Sci.* **48**, S-69–S-84 (2008).

20. Bennett, M. D. & Leitch, I. J. Nuclear DNA amounts in angiosperms: progress, problems and prospects. *Ann. Bot. (Lond.)* **95**, 45–90 (2005).

21. Vogel, J. P. *et al.* EST sequencing and phylogenetic analysis of the model grass *Brachypodium distachyon*. *Theor. Appl. Genet.* **113**, 186–195 (2006).

22. Rajagopalan, R., Vaucheret, H., Trejo, J. & Bartel, D. P. A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis* thaliana. *Genes Dev.* **20**, 3407–3425 (2006).

23. Tanaka, T. *et al.* The rice annotation project database (RAP-DB): 2008 update. *Nucleic Acids Res.* **36**, D1028–D1033 (2008).

24. Fox, S., Filichkin, S. & Mockler, T. Applications of ultra-high-throughput sequencing. *Methods Mol. Biol.* **553**, 79–108 (2009).

25. Gray, J. *et al.* A recommendation for naming transcription factor proteins in the grasses. *Plant Physiol.* **149**, 4–6 (2009).

26. Vogel, J. Unique aspects of the grass cell wall. *Curr. Opin. Plant Biol.* **11**, 301–307 (2008).

27. Bennetzen, J. L. & Kellogg, E. A. Do plants have a one-way ticket to genomic obesity? *Plant Cell* **9**, 1509–1514 (1997).

28. Wicker, T. & Keller, B. Genome-wide comparative analysis of *copia* retrotransposons in Triticeae, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. *Genome Res.* **17**, 1072–1081 (2007).

29. Wicker, T. *et al.* Analysis of intraspecies diversity in wheat and barley genomes identifies breakpoints of ancient haplotypes and provides insight into the structure of diploid and hexaploid triticeae gene pools. *Plant Physiol.* **149**, 258–270 (2009).

30. Jiang, N., Bao, Z., Zhang, X., Eddy, S. R. & Wessler, S. R. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431**, 569–573 (2004).

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

32. Grass Phylogeny Working Group. Phylogeny and subfamilial classification of the grasses (Poaceae). *Ann. Mo. Bot. Gard.* **88**, 373–457 (2001).

33. Bossolini, E., Wicker, T., Knobel, P. A. & Keller, B. Comparison of orthologous loci from small grass genomes *Brachypodium* and rice: implications for wheat genomics and grass genome annotation. *Plant J.* **49**, 704–717 (2007).

34. Charles, M. *et al.* Sixty million years in evolution of soft grain trait in grasses: emergence of the softness locus in the common ancestor of *Pooideae* and *Ehrhartoideae*, after their divergence from *Panicoideae*. *Mol. Biol. Evol.* **26**, 1651–1661 (2009).

35. 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).

36. Stein, N. *et al.* A 1,000-loci transcript map of the barley genome: new anchoring points for integrative grass genomics. *Theor. Appl. Genet.* **114**, 823–839 (2007).

37. Luo, M. C. *et al.* Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in Triticeae. *Proc. Natl Acad. Sci. USA* **106**, 15780–15785 (2009).

38. Qi, L. L. *et al.* A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. *Genetics* **168**, 701–712 (2004).

39. Salse, J. *et al.* Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* **20**, 11–24 (2008).

40. Srinivasachary, Dida M. M., Gale, M. D. & Devos, K. M. Comparative analyses reveal high levels of conserved colinearity between the finger millet and rice genomes. *Theor. Appl. Genet.* **115**, 489–499 (2007).

41. Vicient, C. M., Kalendar, R. & Schulman, A. H. Variability, recombination, and mosaic evolution of the barley BARE-1 retrotransposon. *J. Mol. Evol.* **61**, 275–291 (2005).

42. Meyers, B. C., Kozik, A., Griego, A., Kuang, H. & Michelmore, R. W. Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis. Plant Cell* **15**, 809–834 (2003).

43. Ma, J. & Bennetzen, J. L. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl Acad. Sci. USA* **101**, 12404–12410 (2004).

44. U.S. Department of Energy Office of Science. *Breaking the Biological Barriers to Cellulosic Ethanol: A Joint Research Agenda* 〈http://genomicscience.energy.gov/biofuels/b2bworkshop.shtml〉 (2006).

45. Food and Agriculture Organization of the United Nations. *World Agriculture: Towards 2030/2050 Interim Report*. 〈http://www.fao.org/ES/esd/AT2050web.pdf〉 (2006).

46. McCarthy, E. M. & McDonald, J. F. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* **19**, 362–367 (2003).

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

Acknowledgements We acknowledge the contributions of the late M. Gale, who identified the importance of conserved gene order in grass genomes. This work was mainly supported by the US Department of Energy Joint Genome Institute Community Sequencing Program project with J.P.V., D.F.G., T.C.M. and M.W.B., a BBSRC grant to M.W.B., an EU Contract Agronomics grant to M.W.B. and K.F.X.M., and GABI Barlex grant to K.F.X.M. Illumina transcriptome sequencing was supported by a DOE Plant Feedstock Genomics for Bioenergy grant and an Oregon State Agricultural Research Foundation grant to T.C.M.; small RNA research was supported by the DOE Plant Feedstock Genomics for Bioenergy grants to P.J.G. and T.C.M.; annotation was supported by a DOE Plant Feedstocks for Genomics Bioenergy grant to J.P.V. A full list of support and acknowledgements is in the Supplementary Information.

Author Information The whole-genome shotgun sequence of *Brachypodium distachyon* has been deposited at DDBJ/EMBL/GenBank under the accession ADDN00000000. (The version described in this manuscript is the first version, accession ADDN01000000). EST sequences have been deposited with dbEST (accessions 67946317–68053959) and GenBank (accessions GT758162–GT865804). The short read archive accession for RNA-seq data is SRA010177. Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.P.V. (john.vogel@ars.usda.gov) or D.F.G. (david.garvin@ars.usda.gov) or T.C.M. (tmockler@cgrb.oregonstate.edu) or M.W.B. (michael.bevan@bbsrc.ac.uk).

Author Contributions See list of consortium authors below.

The International Brachypodium Initiative

Principal investigators John P. Vogel[1], David F. Garvin[2], Todd C. Mockler[3], Jeremy Schmutz[4], Dan Rokhsar[5,6], Michael W. Bevan[7]; **DNA sequencing and assembly** Kerrie Barry[5], Susan Lucas[5], Miranda Harmon-Smith[5], Kathleen Lail[5], Hope Tice[5], Jeremy Schmutz[4] (Leader), Jane Grimwood[4], Neil McKenzie[7], Michael W. Bevan[7]; **Pseudomolecule assembly and BAC end sequencing** Naxin Huo[1], Yong Q. Gu[1], Gerard R. Lazo[1], Olin D. Anderson[1], John P. Vogel[1] (Leader), Frank M. You[8], Ming-Cheng Luo[8], Jan Dvorak[8], Jonathan Wright[7], Melanie Febrer[7], Michael W. Bevan[7], Dominika Idziak[9], Robert Hasterok[9], David F. Garvin[2]; **Transcriptome sequencing and analysis** Erika Lindquist[5], Mei Wang[5], Samuel E. Fox[3], Henry D. Priest[3], Sergei A. Filichkin[3], Scott A. Givan[3], Douglas W. Bryant[3], Jeff H. Chang[3], Todd C. Mockler[3] (Leader), Haiyan Wu[10,24], Wei Wu[10], An-Ping Hsia[10], Patrick S. Schnable[10,24], Anantharaman Kalyanaraman[11], Brad Barbazuk[12], Todd P. Michael[13], Samuel P. Hazen[14], Jennifer N. Bragg[1], Debbie Laudencia-Chingcuanco[1], John P. Vogel[1], David F. Garvin[2], Yiqun Weng[15], Neil McKenzie[7], Michael W. Bevan[7]; **Gene analysis and annotation** Georg Haberer[16], Manuel Spannagl[16], Klaus Mayer[16] (Leader), Thomas Rattei[17], Therese Mitros[6], Dan Rokhsar[6], Sang-Jik Lee[18], Jocelyn K. C. Rose[18], Lukas A. Mueller[19], Thomas L. York[19]; **Repeats analysis** Thomas Wicker[20] (Leader), Jan P. Buchmann[20], Jaakko Tanskanen[21], Alan H. Schulman[21] (Leader), Heidrun Gundlach[16], Jonathan Wright[7], Michael Bevan[7], Antonio Costa de Oliveira[22], Luciano da C. Maia[22], William Belknap[1], Yong Q. Gu[1], Ning Jiang[23], Jinsheng Lai[24], Liucun Zhu[25], Jianxin Ma[25], Cheng Sun[26], Ellen Pritham[26]; **Comparative genomics** Jerome Salse[27] (Leader), Florent Murat[27], Michael Abrouk[27], Georg Haberer[16], Manuel Spannagl[16], Klaus Mayer[16], Remy Bruggmann[13], Joachim Messing[13], Frank M. You[8], Ming-Cheng Luo[8], Jan Dvorak[8]; **Small RNA analysis** Noah Fahlgren[3], Samuel E. Fox[3], Christopher M. Sullivan[3], Todd C. Mockler[3], James C. Carrington[3], Elisabeth J. Chapman[3,28], Greg D. May[29], Jixian Zhai[30], Matthias Ganssmann[30], Sai Guna Ranjan Gurazada[30], Marcelo German[30], Blake C. Meyers[30], Pamela J. Green[30] (Leader); **Manual annotation and gene family analysis** Jennifer N. Bragg[1], Ludmila Tyler[1,6], Jiajie Wu[1,8], Yong Q. Gu[1], Gerard R. Lazo[1], Debbie Laudencia-Chingcuanco[1], James Thomson[1], John P. Vogel[1] (Leader), Samuel P. Hazen[14], Shan Chen[14], Henrik V. Scheller[31], Jesper Harholt[32], Peter Ulvskov[32], Samuel E. Fox[3], Sergei A. Filichkin[3], Noah Fahlgren[3], Jeffrey A. Kimbrel[3], Jeff H. Chang[3], Christopher M. Sullivan[3], Elisabeth J. Chapman[3,27], James C. Carrington[3], Todd C. Mockler[3], Laura E. Bartley[8,31], Peijian Cao[8,31], Ki-Hong Jung[8,31]†, Manoj K Sharma[8,31], Miguel Vega-Sanchez[8,31], Pamela Ronald[8,31], Christopher D. Dardick[33], Stefanie De Bodt[34], Wim Verelst[34], Dirk Inzé[34], Maren Heese[35], Arp Schnittger[35], Xiaohan Yang[36], Udaya C. Kalluri[36], Gerald A. Tuskan[36], Zhihua Hua[37], Richard D. Vierstra[37], David F. Garvin[3], Yu Cui[24], Shuhong Ouyang[24], Qixin Sun[24], Zhiyong Liu[24], Alper Yilmaz[38], Erich Grotewold[38], Richard Sibout[39], Kian Hematy[39], Gregory Mouille[39], Herman Höfte[39], Todd Michael[13], Jérome Pelloux[40], Devin O'Connor[41], James Schnable[41], Scott Rowe[41], Frank Harmon[41], Cynthia L. Cass[42], John C. Sedbrook[42], Mary E. Byrne[7], Sean Walsh[7], Janet Higgins[7], Michael Bevan[7], Pinghua Li[19], Thomas Brutnell[19], Turgay Unver[43], Hikmet Budak[43], Harry Belcram[44], Mathieu Charles[44], Boulos Chalhoub[44], Ivan Baxter[45]

[1]USDA-ARS Western Regional Research Center, Albany, California 94710, USA. [2]USDA-ARS Plant Science Research Unit and University of Minnesota, St Paul, Minnesota 55108, USA. [3]Oregon State University, Corvallis, Oregon 97331-4501, USA. [4]HudsonAlpha Institute, Huntsville, Alabama 35806, USA. [5]US DOE Joint Genome Institute, Walnut Creek, California 94598, USA. [6]University of California Berkeley, Berkeley, California 94720, USA. [7]John Innes Centre, Norwich NR4 7UJ, UK. [8]University of California Davis, Davis, California 95616, USA. [9]University of Silesia, 40-032 Katowice, Poland. [10]Iowa State University, Ames, Iowa 50011, USA. [11]Washington State University, Pullman, Washington 99163, USA. [12]University of Florida, Gainsville, Florida 32611, USA. [13]Rutgers University, Piscataway, New Jersey 08855-0759, USA. [14]University of Massachusetts, Amherst, Massachusetts 01003-9292, USA. [15]USDA-ARS Vegetable Crops Research Unit, Horticulture Department, University of Wisconsin, Madison, Wisconsin 53706, USA. [16]Helmholtz Zentrum München, D-85764 Neuherberg, Germany. [17]Technical University München, 80333 München, Germany. [18]Cornell University, Ithaca, New York 14853, USA. [19]Boyce Thompson Institute for Plant Research, Ithaca, New York 14853-1801, USA. [20]University of Zurich, 8008 Zurich, Switzerland. [21]MTT Agrifood Research and University of Helsinki, FIN-00014 Helsinki, Finland. [22]Federal University of Pelotas, Pelotas, 96001-970, RS, Brazil. [23]Michigan State University, East Lansing, Michigan 48824, USA. [24]China Agricultural University, Beijing 10094, China. [25]Purdue University, West Lafayette, Indiana 47907, USA. [26]The University of Texas, Arlington, Arlington, Texas 76019, USA. [27]Institut National de la Recherché Agronomique UMR 1095, 63100 Clermont-Ferrand, France. [28]University of California San Diego, La Jolla, California 92093, USA. [29]National Centre for Genome Resources, Santa Fe, New Mexico 87505, USA. [30]University of Delaware, Newark, Delaware 19716, USA. [31]Joint Bioenergy Institute, Emeryville, California 94720, USA. [32]University of Copenhagen, Frederiksberg DK-1871, Denmark. [33]USDA-ARS Appalachian Fruit Research Station, Kearneysville, West Virginia 25430, USA. [34]VIB Department of Plant Systems Biology, VIB and Department of Plant Biotechnology and Genetics, Ghent University, Technologiepark 927, 9052 Gent, Belgium. [35]Institut de Biologie Moléculaire des Plantes du CNRS, Strasbourg 67084, France. [36]BioEnergy Science Center and Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6422, USA. [37]University of Wisconsin-Madison, Madison, Wisconsin 53706, USA. [38]The Ohio State University, Columbus, Ohio 43210, USA. [39]Institut Jean-Pierre Bourgin, UMR1318, Institut National de la Recherche Agronomique, 78026 Versailles cedex, France. [40]Université de Picardie, Amiens 80039, France. [41]Plant Gene Expression Center, University of California Berkeley, Albany, California 94710, USA. [42]Illinois State University and DOE Great Lakes Bioenergy Research Center, Normal, Illinois 61790, USA. [43]Sabanci University, Istanbul 34956, Turkey. [44]Unité de Recherche en Génomique Végétale: URGV (INRA-CNRS-UEVE), Evry 91057, France. [45]USDA-ARS/Donald Danforth Plant Science Center, St Louis, Missouri 63130, USA. †Present address: The School of Plant Molecular Systems Biotechnology, Kyung Hee University, Yongin 446-701, Korea.

Supplementary Information

Genome Sequence and Assembly

Nuclear DNA was prepared from *Brachypodium distachyon* (*Brachypodium*) Bd21 plants derived by single- seed descent for 8 generations to reduce potential sequence polymorphism. Plants were grown at 20°C in a greenhouse in long day conditions for 3 weeks and transferred to darkness for 2 days prior to nuclei isolation to reduce starch levels. Nuclei were prepared [1] with an additional Percoll gradient purification of nuclei. High molecular weight DNA was extracted and purified by gentle lysis, phenol/$CHCl_3$ extraction and dialysis. Libraries were prepared from nuclear DNA (Supplementary Table 1) and sequenced using standard Sanger protocols on ABI 3730 xl instruments. The total number of reads from each library is shown in Supplementary Table 1.

Supplementary Table 1. Assembly input. The whole genome shotgun strategy involved end-sequencing different sized insert libraries. These are shown below, together with their mean insert size, number of reads from each library, and estimated genome coverage.

Library	Insert Size	Reads	Coverage
3kb (1)	3,215	277,248	0.65
3kb (2)	3,237	1,519,924	3.17
8kb (1)	6,381	855,422	2.04
8kb (2)	6,392	1,448,347	2.46
fosmid (1)	32,823	60,767	0.06
fosmid (2)	35,691	325,536	0.52
BAC BRA (BAC DH)	94,073	110,592	0.22
BAC BRB (BAC DB)	101,562	36,864	0.08
BAC DH [1] (HinDIII)	103,216	30,704	0.05
BAC DB [1] (BamH1)	108,177	36,388	0.04
BAC BD_CBa [2] (EcoR1)	124,935	25,948	0.05
BAC BD_ABa [2] (HinDIII)	149,112	34,177	0.07
TOTAL		4,761,917	9.43

[1] BAC libraries DH and DB are described in [2-4]. Details of BAC libraries BD_CBa and BD_ABa will be published elsewhere.

Construction of the preliminary scaffold assemblies.

A total of 4,761,917 reads (see Supplementary Table 1 for clone sizes) were assembled with a modified version of Arachne[5] v.20071016 with parameters maxcliq1=100, correct1_passes=0 and BINGE_AND_PURGE=True to form 217 scaffolds covering 272.1 Mb of the *Brachypodium* genome (see Supplementary Table 2 for scaffold and contigs totals).

Supplementary Table 2. Raw assembly output. Summary statistics of the output of the whole genome shotgun assembly, before breaking and constructing chromosome scale assemblies and before contamination based screening. Total contigs and total assembled base-pairs for each set of scaffolds greater than the given size are also shown.

Scaffold Length (bp)	Number of Scaffolds	Number of Contigs	Total Scaffold Length (bp)	Total Contig Length (bp)	Coverage
all	217	2,067	272,287,606	272,077,374	99.66%
>5,000	127	1,925	272,020,434	271,781,248	99.66%
>50,000	13	1,684	270,814,201	270,471,535	99.65%
>500,000	11	1,671	270,737,212	270,363,712	99.61%
>5,000,000	10	1,665	270,190,573	269,833,561	99.60%

Generation of *Brachypodium* v1.0 pseudomolecules and final assembly.

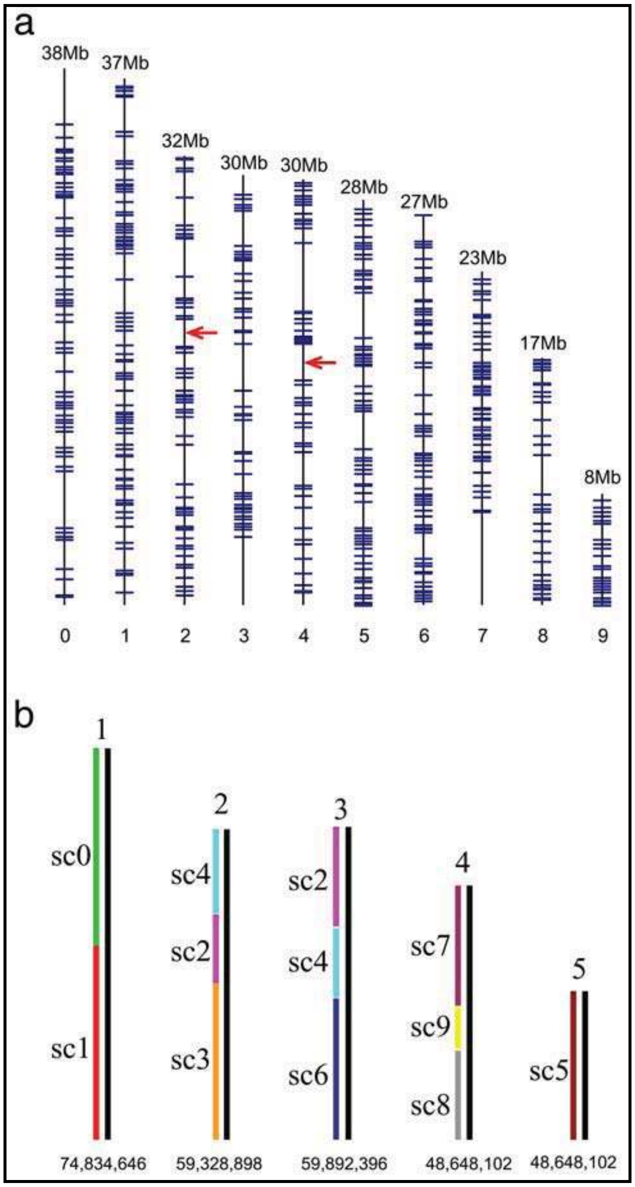
Based on the genetic map integration two breaks were made in the scaffolds (Supplementary Figure 1). The scaffolds were then ordered with Arachne, making 7 map based joins to create 5 chromosome-scale pseudomolecules. Each scaffold was oriented and joined with 10,000 N bps to signify a map join. The scaffolds were then compared again with the genetic map to verify the ordering, and assigned to pseudomolecules 1-5 according to the karyotype (Supplementary Figure 2) and *Brachypodium* genetic linkage groups. The remaining scaffolds were classified depending on sequence content. Contamination was detected using megablast against Genbank NR and blastp against a set of known microbial proteins. No prokaryotic contamination was identified. Scaffolds not included in the final assembly were: unanchored rDNA (51); mitochondrial (2); chloroplastic (14); and small unanchored repetitive scaffolds as defined by 95% of the 24mers occurring greater than four times in the large scaffolds (43) or were less than 1kb in sequence length (2). We appended the remaining 78 scaffolds to the 5 chromosome scaffolds. The resulting final genome assembly statistics are shown in Supplementary Table 3.

Supplementary Table 3. Final summary assembly statistics for chromosome scale assembly.

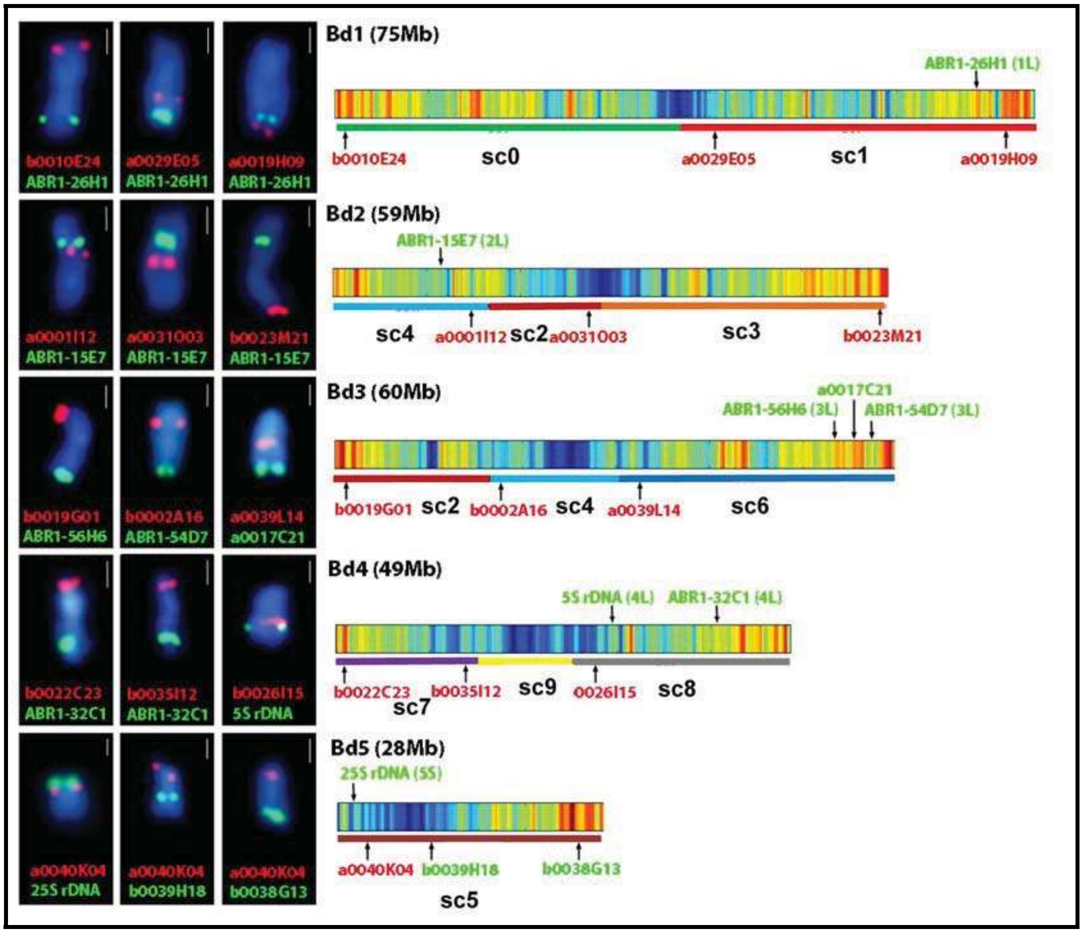
Final Contigs	1,754
Total Genome Size	271,148,425 bp
Mapped Sequence Size	270,058,955 bp
Estimated Gaps	1,089,470 bp (0.4% of genome)
Release Scaffold Total	83 (50<10 Kb)
Release Contig Total	1,754
Release Scaffold Sequence Total	271.9 Mb
Release Contig Sequence Total (estimate 0.4% gaps)	270.8 Mb
Release Scaffold N/L50	3/59.3 Mb
Release Contig N/L50	252/347.8 Kb
Number of scaffolds >50KB	6
Final Genome Coverage	9.4x

Organelle DNA in the nuclear genome

A total of 1,131 chloroplast DNA insertions covering 275,328 bp (0.10%) of the nuclear genome, and 2,107 insertions of mitochondrial DNA covering 487,793 bp (0.18%) of the nuclear genome were found. Most insertions were less than 0.5 kb, but 17 chloroplast insertions contained intact genes, and approximately 23% of chloroplast and 8% of mitochondrial insertions were identical to organelle sequences, indicating ongoing insertion events.



Supplementary Figure 1. Ordering sequence scaffolds using a genetic map. To verify and assemble the 10 largest preliminary scaffolds (sc0-sc9) into chromosome-scale assemblies we compared the scaffolds to a high-density genetic map constructed from 562 SNP markers selected to be evenly spaced along the scaffolds (full details of the map will be published elsewhere). (a) The locations of genetic markers on the scaffolds are indicated by blue lines. Only two false joins were detected and scaffolds two and four were broken where indicated by red arrows. Scaffold number is indicated below and scaffold length is indicated on the top of each scaffold. (b) Color coded assignment of scaffolds to the five *Brachypodium* chromosomes. Chromosome number is indicated above and total length in bp is indicated below each chromosome.



Supplementary Figure 2. Aligning genome sequence assemblies to *Brachypodium* chromosomes. Scaffolds (sc0-9) from the sequence assemblies were aligned to the *Brachypodium* karyotype using fluorescently labelled BACs from a physical map integrated with the sequence assemblies (MF, JW, MWB, in preparation). The methods used are described below. Reference BACs with known chromosomal locations (ABR1 clones) and 5S rDNA and 25S rDNA markers, shown in green, are from[6]. Red (or green, clones a007C21, b0039H18 and b0038G13) fluorescence shows the position of individual BACs integrated into the sequence scaffolds identified as lines under the pseudomolecule heatmaps showing gene density. The scale bar in the micrographs is 1μm. The size of each chromosome is shown and the scaffolds are coloured according to Supplementary Figure 1.

In situ hybridization

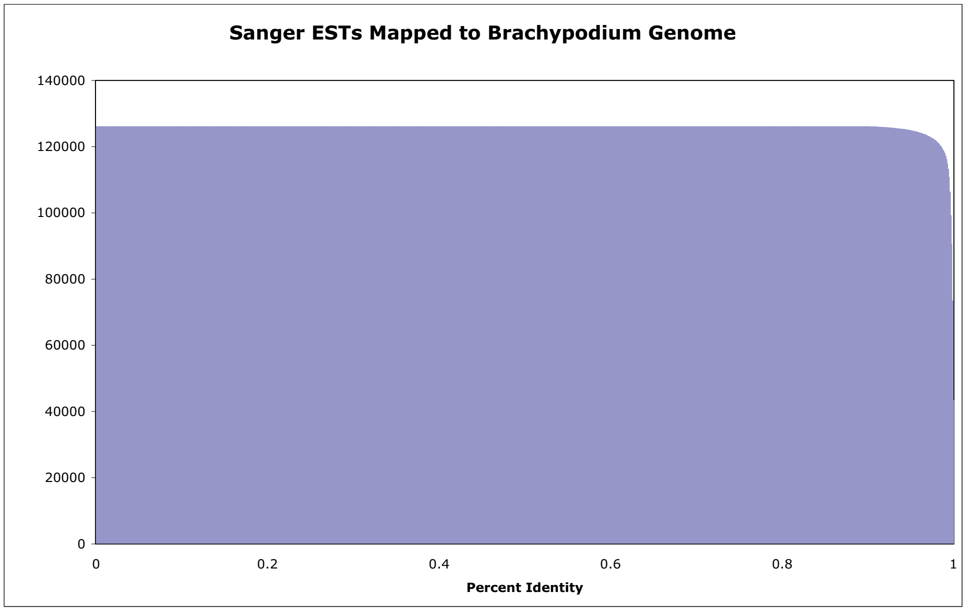
Metaphase chromosome spreads were made from excised and fixed *Brachypodium* Bd21 roots grown for 3-5 days, essentially as described [7]. BACs were identified for labelling from a physical map of *Brachypodium* (MF, JW. MWB, in preparation) that was integrated with genome sequence assemblies. Reference BACs with known chromosomal locations [6] were selected from the ABR1/ABR5 libraries. Isolated BAC DNA was labelled by nick-translation with digoxigenin-11-dUTP (Roche) or tetramethyl-rhodamine-5-dUTP. A 2.3-kb *Cla*I subclone of the 25S rDNA coding region of *A. thaliana* [8] was used to visualize the 45S rDNA locus that is diagnostic for short arm of chromosome 5. A 5S rDNA probe was obtained from the wheat clone pTa794 [9] by PCR amplification. This probe was used to visualise the 5S rDNA locus, diagnostic for long arm of chromosome 4. The general conditions of FISH procedure were as follows: the high-stringency (77% sequence identity) hybridization mixture was 50% deionized formamide, 20% dextran sulfate, 2x SSC and salmon sperm blocking DNA in 25-100x excess of labelled probes. All probes were mixed to a final concentration each of 2 - 5 ng/μl of the mixture and denatured (75 ºC for 10 min). The slides with chromosome material and the hybridization mixture were then denatured together for 4.5 min at 70 ºC and allowed to hybridise for 12-20 h in a humid chamber at 37 ºC. Post-hybridisation washes were carried out for 10 min in 10% deionised formamide in 0.1× SSC at 42 ºC. Digoxigenated probes were immunodetected using standard protocol for FITC-conjugated anti-digoxigenin antibodies (Roche) and visualized as green fluorescence signals. The preparations were mounted and counterstained in Vectashield containing 2.5 μg/ml of 4',6-diamidino-2-phenylindole (DAPI; Serva).

urces used for genome annotation. The table describes the tissues used as sources of RNA for

quenced by	Bd genotype	Tissue/Stage/Treatment etc…	Normalization	Contributor/Reference
l	Bd21	callus	N/A	Vogel, Bragg
l	Bd21	roots	DSN	Garvin
l	Bd21	developing seeds	DSN	Mockler, Michael, Laudencia-Chingcuanco
l	Bd21	diurnally sampled whole seedlings	DSN	Mockler
l	Bd21	diurnally sampled roots	DSN	Garvin
l	Bd21	diurnally sampled leaves + stems	DSN	Mockler
l	Bd21	diurnally sampled flowers RNA	DSN	Mockler
l	Bd21	callus	DSN	Vogel, Bragg
l	Bd21	diurnally sampled leaves + stems + callus	DSN	Mockler, Vogel, Bragg
l	Bd21	diurnally sampled leaves + stems + callus	DSN	Mockler, Vogel, Bragg
l	Bd21	diurnally sampled leaves + stems + callus	DSN	Mockler, Vogel, Bragg
hnable	PI 185133 (source of Bd2-3)	root tips	N/A	Schnable
hnable	PI 185134 (source of Bd3-1 and 3-2)	root tips	N/A	Schnable
hnable	PI 245730 (source of Bd18-1)	root tips	N/A	Schnable
hnable	PI 254867 (source of Bd21)	root tips	N/A	Schnable
l	Bd21	abiotic stress + biotic stress	DSN	Mockler, Chang, Hazen, Weng
l	Bd21	superpool	DSN	Mockler, Vogel, Hazen, Chang, Michael, Garvin, Bevan
l	Bd21	flower + flower drought	DSN	Bevan
l	Bd21	leaf+ leaf drought	DSN	Bevan
gel	Bd21	callus	N/A	Vogel
gel	Bd21	leaf	N/A	Vogel
gel	Bd21	root	N/A	Vogel
gel	Bd21	seed	N/A	Vogel
gel	Bd21	stem	N/A	Vogel
ckler	Bd21	superpool	DSN	Mockler, Vogel, Hazen, Chang, Michael, Garvin, Bevan, Laudencia-Chingcuanco, Weng

Illumina Transcriptome Methods

Full-length enriched (FL) and randomly primed (RP) cDNA libraries were prepared from RNA isolated as described in Supplementary Table 4, and sequenced using an Illumina 1G Genome Analyzer essentially as described [10]. Raw Illumina reads were obtained after base calling in the Solexa Pipeline version 0.2.2.6. We removed Illumina reads matching SMART adapters, Solexa sequencing adapters and reads of low quality (containing ambiguous nucleotide calls), and then the low quality bases at the 3' ends of reads were trimmed. Reads were truncated to the first 32 bases. The *Brachypodium* v1.0 genome annotation and Perl scripts were used to generate sequence files representing annotated genome features (exons, introns, UTRs, genes, splice junctions, cDNAs, CDS). Perfect match 32-mer Illumina reads were mapped to the *Brachypodium* v1.0 annotated genome features using HashMatch (http://mocklerlab-tools.cgrb.oregonstate.edu/). Illumina read coverage along the predicted sequence features was calculated using a Perl script to process HashMatch alignment data for each type of sequence feature. Illumina coverage was calculated as the percentage of bases along the length of the sequence feature that were independently supported by Illumina reads. For validation of predicted alternative splicing events, database queries were used to identify all possible "informative" 32-mers unique to specific predicted alternative splice variants among the Bradi v1.0 gene models. Alternative splicing events were validated using a Perl script to match Illumina transcript reads to the database of informative 32-mers representing specific predicted alternative splice variants.



Supplementary Figure 3. Mapping *Brachypodium* Sanger ESTs onto the genomic sequence. *Brachypodium* Sanger ESTs were anchored onto the genomic assemblies as spliced alignments using BLAT. In total, 126,072 out of 128,221 transcript sequences (98.3%) could be mapped to the genomic sequence with a minimum alignment length of 50 nucleotides. On the y-axis, the cumulative frequency of anchored ESTs is shown according to its dependence of alignment identity on the x-axis. In cases where an EST matched several genomic positions, the highest alignment identity was selected. The large majority of ESTs could be mapped with high sequence identities, ≥124,876 (97.4%) and ≥126,072 (98.3%) sequences with an identity ≥95% and ≥90%, respectively.

Small RNA library construction and sequencing.

Brachypodium Bd21 was used for the preparation of two panicle (flower) libraries. For library OBD01, plants were grown in long-day conditions (16 h days/8 h nights) at 25^{o}C. Inflorescence tissue was collected (day 28-35) at 4 time point intervals of 0700 (dawn), 1300, 1900, 0100 hours, and frozen immediately in liquid nitrogen. Tissues were ground in liquid nitrogen and placed at -80°C. For BDI05, panicle tissue was harvested from plants grown at 20°C in 20 h light/4 h dark cycles for 6 weeks. Emerging panicles, excluding flag leaves, were harvested at approximately 10 h into the subjective day. Light intensity for both OBD01 and BDI05 was approximately 120-140 umol m^{-2} sec^{-1}. OBD01 total RNA was extracted using Trizol reagent (Invitrogen) as described in [11] with the following modifications. Equal amounts of tissues from each of the 4 time points were pooled together. The tissue samples were homogenized with Trizol reagent (10 [v/w]) and incubated for 5 minutes at room temperature. Plant debris was separated by centrifugation, and the soluble fraction was extracted three times with chloroform (0.2 [v/v]). Total RNA was precipitated with cold isopropanol and pelleted by centrifugation at 8,400 x g for 30 minutes at 4°C. The RNA pellet was resuspended in 0.1X TE. Small RNA libraries were prepared as previously described in [12] with modifications. Throughout small RNA isolation and adaptor ligation steps, RNA samples were size-selected by gel electrophoresis as follows. RNA was denatured for 4 minutes at 100°C and resolved by electrophoresis on 17% polyacrylamide gels containing 7 M urea in 0.5X TBE buffer (45 mM Tris-borate, pH 8.0, and 1.0 mM EDTA). Gel slices containing RNA that co-migrated with ^{32}P-radiolabeled size standards were excised. RNA was electrophoretically transferred to DE81 chromatography paper (Fisher Scientific) and recovered by incubation at 70°C in high salt buffer (10 mM Tris-HCl, pH 7.6; 1 mM EDTA; 1 M NaCl; 50 mM L-Arginine) followed by ethanol precipitation with glycogen (20 µg) for 4 hours at -80°C. Ligation of the 3' adaptor (miRNA cloning linker-1, 5'-rAppTGGAATTCTCGGGTGCCAAGG/ddC/-3'; IDT) to 18 - 24 nt RNA was done by 12 hour incubation at 4°C with T4 RNA ligase (Ambion). Following size selection, RNA was ligated to the 5' RNA oligonucleotide adaptor (5'-GUUCAGAGUUCUACAGUCCGACGAUC-3') and size-selected as described above. Following reverse transcription and second strand synthesis (RT-primer, 5'-ATTGATGGTGCCTACAG-3'), cDNA was amplified by 26 cycles of PCR using Phusion High-Fidelity DNA Polymerase (New England Biolabs). The 5' PCR primer (5'-AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA-3'), and 3' PCR primer (5'-CAAGCAGAAGACGGCATACGAATTGATGGTGCCTACAG-3') contained sequences required for cluster generation on the Illumina Genome Analyzer system. DNA amplicons (2.5 pmol) were added to each flow-cell lane following the Illumina protocol (Illumina, http://www.illumina.com). The library was sequenced (36 cycles; sequencing primer, 5'-GTTCAGAGTTCTACAGTCCGA-3') using an Illumina Genome Analyzer at the Center for Genome Research and Biocomputing at Oregon State University. Similarly, for BDI05 panicle tissues, total RNA was isolated using Trizol reagent and small RNA libraries were constructed according to [13,14]. The 5' RNA adapter was 5' GUUCAGAGUUCUACAGUCCGACGAUC 3' and the RNA 3' adapter was 5' P-UCGUAUGCCGUCUUCUGCUUG-idT 3'. The forward PCR primer was 5' AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA 3' and the reverse PCR primer was 5' CAAGCAGAAGACGGCATACGA 3'. The library was sequenced (36 cycles; sequencing primer, 5' CGACAGGTTCAGAGTTCTACAGTCCGACGATC 3') using an Illumina Genome Analyzer at the National Center for Genome Resources.

Analysis of phased small RNAs.

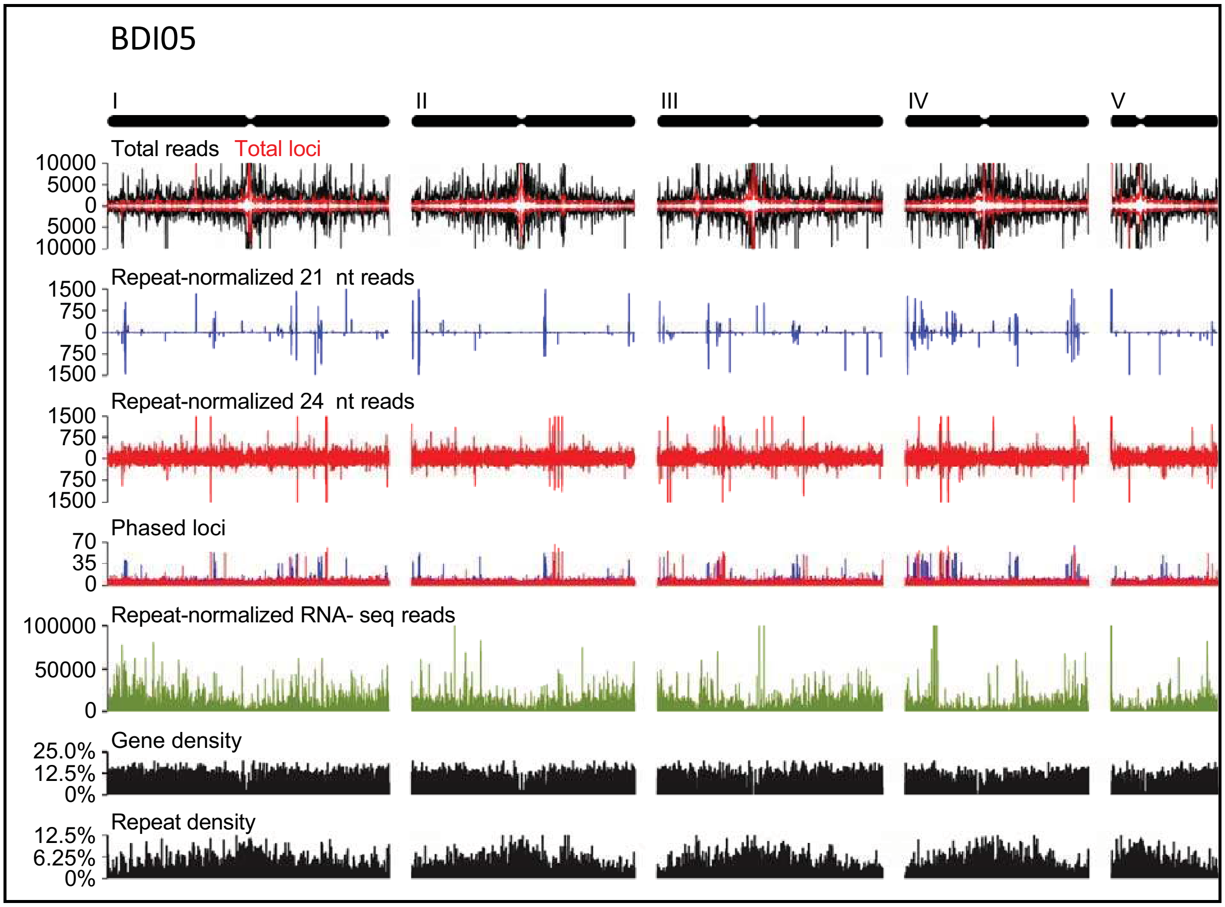
To identify genomic regions generating phased small RNAs, we modified an algorithm designed for 454 data [15], adapting it to the higher sequencing depth produced by SBS sequencing. Phasing scores were assigned to each 10-cycle window, based on the following formula:

Phasing score = $$\ln\left[\left(1 + 10 \times \frac{\sum_{i=1}^{10} \mathrm{Pi}}{1 + \sum \mathrm{U}}\right)^{n-2}\right]$$, n > 3

n: number of phase cycle positions occupied by at least one small RNA (allowing a shift of plus or minus one nucleotide) within a ten-cycle window.
P: the total number of reads for all small RNAs with start coordinates in a given phase (allowing a shift of plus or minus one nucleotide) within a ten-cycle window.
U: the total number of reads for all small RNAs with start coordinates out of the given phase within the ten-cycle window.

In this analysis, the abundance of each position is calculated as the sum of abundances of all small RNAs from the sense strand sharing the same 5' starting position, summed with the abundance of small RNAs from the anti-sense strand that form a complementary pair (a duplex with a two nucleotides 3'-overhang). The calculation of abundance was essentially as described previously [15]. In addition, if the highest abundance at any one position comprised more than 90% of the total abundance in the entire ten-cycle window, this position was omitted, to avoid including highly abundance miRNA loci.

This method was applied to the *Brachypodium* small RNA libraries, which identified the highest numbers of phased clusters in the inflorescence libraries, and these were used for further analysis. As a comparison, the same algorithm was also applied to a published, wild-type Arabidopsis inflorescence library available in GenBank's GEO as GSM284747.



Supplementary Figure 4. Genome-wide distribution of small RNA genes identified in the BDI05 panicle library and their alignment with repeat elements in the *Brachypodium* genome. Each of the five *Brachypodium* chromosomes are shown as ideograms at the top of each figure. Total reads and total loci graphs plot total small RNA reads (black lines) and total small RNA loci (red lines). Repeat-normalized 21 nt reads and repeat-normalized 24 nt reads histograms plot 21 or 24 nt small RNA reads normalized for repeated matches to the genome, respectively. Phased loci histograms plot the position and phase-score of 21 (blue) and 24 (red) nt phased small RNA loci. Repeat-normalized RNA-seq reads histograms plot the abundance of reads matching RNA transcripts, normalized for ambiguous matches to the genome. Gene and repeat density histograms plot the percentage of nucleotide space occupied by genes (exons + introns) or repeats (transposons, retrotransposons and centromeric repeats). Plots for total small RNA reads, total small RNA loci, repeat-normalized 21 and 24 nt small RNA reads, repeat-normalized RNA-seq reads, gene density and repeat density were generated using the scrolling window method

(window = 100,000 nt, scroll = 20,000 nt).

or analysis of small RNA phasing intervals in the *Brachypodium* genome. Gray regions of table articular interest, exceeding an arbitrary cut-off score of 25. “Position number” indicates the number of above a specific score, “cluster number” indicates the number of loci at or above the score; all high dow were combined to generate one cluster.

20		21		22		23		24		25	
position number	cluster number	position number	cluster number	position number	cluster number	position number	cluster number	position number	cluster number	position number	cluster number
22,985	2,295	18,696	2,082	14,607	1,786	12,049	1,661	10,386	1,545	9,386	1,398
2,962	679	2,343	537	1,696	426	1,251	342	1,118	330	918	278
416	118	401	91	260	78	175	35	153	36	132	46
75	19	182	26	66	7	84	4	73	5	49	12
33	4	100	17	39	5	43	2	26	3	18	5
13	3	53	13	12	2	14	3	7	3	4	1
5	2	29	7	0	0	2	2	0	0	0	0
0	0	21	5	0	0	0	0	0	0	0	0
0	0	7	4	0	0	0	0	0	0	0	0
0	0	2	2	0	0	0	0	0	0	0	0
10,177	3,073	16,421	3,517	7,399	2,392	6,537	2,160	11,196	2,254	5,327	1,766
2,616	750	9,085	1,551	1,749	538	1,566	452	6,217	748	1,399	398
801	201	6,587	1,074	497	144	449	113	4,635	393	516	120
271	65	5,140	838	160	43	135	45	3,882	299	189	46
81	25	4,083	693	51	18	32	10	3,414	257	30	15
17	8	3,224	589	18	10	2	2	3,056	227	10	5
6	5	2,519	509	2	2	0	0	2,756	213	0	0
2	2	1,865	413	1	1	0	0	2,462	198	0	0
0	0	1,348	329	1	1	0	0	2,203	188	0	0
0	0	951	252	0	0	0	0	1,924	180	0	0
11,767	2,671	18,887	3,302	8,661	2,209	7,625	1,906	11,787	2,077	6,094	1,537
2,846	708	10,410	1,592	1,852	558	1,701	487	5,893	749	1,435	358
683	205	7,687	1,146	384	144	377	134	4,353	409	325	96
173	61	6,387	986	118	56	132	36	3,750	303	114	34
55	26	5,355	877	43	20	59	20	3,404	254	61	11
24	14	4,472	776	17	12	32	12	3,088	235	40	8
4	3	3,625	668	10	7	24	7	2,797	227	31	6
0	0	2,876	579	4	2	14	5	2,504	217	16	6
0	0	2,236	473	1	1	11	5	2,240	210	9	5
0	0	1,661	386	0	0	8	5	1,976	190	6	5

cleotides between small RNAs, analyzed in a 10-phase window across the genome. The algorithm
plementary information above.
was previously described [16].

Protein-coding and tRNA gene predictions

Protein coding gene models were derived from weighted consensus predictions based on several types of evidence: *ab initio* gene finders, protein homology and optimal spliced alignments of expressed sequence tags (ESTs) and tentative consensus transcripts (TCs). Gene finders included the programs Fgenesh++ and Protmap using the monocot Markov models and the Uniref database, GeneID using the wheat Markov models and the PASA pipeline applying Fgenesh predictions and transcripts of *Brachypodium,* wheat and barley. All ESTs, transcript assemblies and reference proteins were mapped as optimal spliced alignments on the whole genome sequence using GenomeThreader [17] and a splice site model of rice. A minimum coding size of 50 amino acids and a minimal spliced mapping size of 50% of the evidence sequence length were required. Intron sizes were constrained to a minimum of 50 bp and a maximum of 30 kb. Protein sets of three finished plant genome projects: rice (version TIGR5 and RAP2) [18,19]; sorghum (version 1.4) [20]; and Arabidopsis (version TAIR8) [21,22] were used to derive protein homologies. Optimal spliced alignments of TIGR transcript assemblies comprising several monocotyledonous species (*Zea mays*, *Saccharum officinale*, *Oryza sativa*, *Hordeum vulgare*, *Triticum aestivum* and *Brachypodium distachyon*) were used for gene predictions based on homology and/or experimental evidence. Supplementary Table 4 describes *Brachypodium* ESTs derived by Sanger and 454 sequencing. This experimental evidence and *ab initio* predictions were used to generate a training set of 410 gene models. The statistical combiner JIGSAW [23] was trained based on this gene set and then applied to the whole genome sequence to integrate experimental evidence into a consensus gene model for each locus. These gene models were rerun through the PASA pipeline to predict UTRs from EST information, to identify possible alternative splicing patterns, and to fit all predicted models to the splice sites supported by EST evidence. Predicted genes were given a unique chromosome location identifier based on the original Arabidopsis convention [24] in which Bradi refers to *Brachypodium distachyon*.

Predicted genes were classified into six confidence classes based on their similarity, size differences, alignment coverage and alignment continuity to proteins in a reference database compiled from SWISSPROT, rice (RAP2 and TIGR5), sorghum (version 1.4) and Arabidopsis (TAIR8) protein databases (Supplementary Figure 5). Protein size differences (coverage) were determined as the quotient of source and reference protein size. Alignment coverage between source and reference protein was defined as twice the alignment length divided by the sum of source and reference protein sizes. Alignment continuity was determined from optimal local Smith-Waterman alignments using the BLOSUM62 similarity matrix and sliding windows of size 10 and overlap of 8 amino acids. It was measured as ratio of alignment slices that contain at least 6 aligned similar amino acids versus the number of aligned 10mers with five or more mismatches or gaps. Gene predictions with no or low homology support (classes 0 and 1, Supplementary Figure 5) were independently evaluated for transcriptional evidence using 10.2 Gb Illumina transcriptome data. Sixty-eight percent of class 0 and 1 models were retired because they had no PASA support or less than 20% coverage over the length of the predicted cDNA by Illumina data (Supplementary Figure 5).

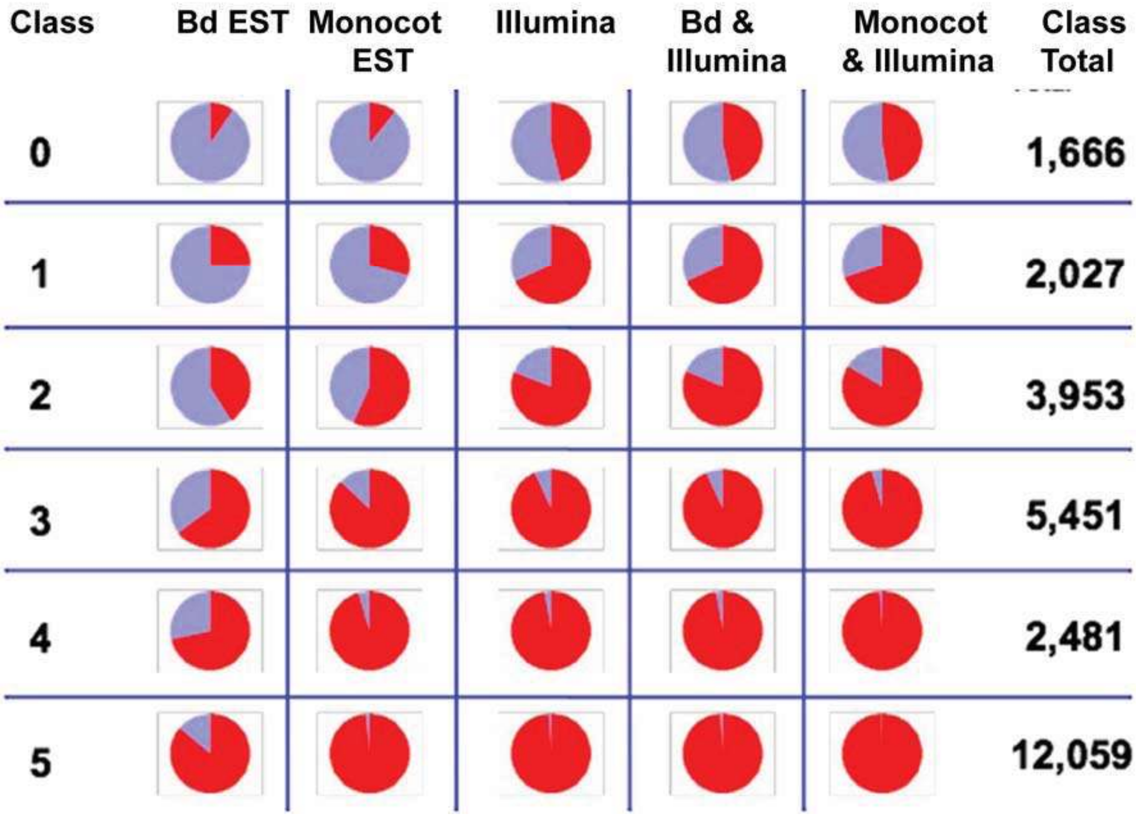
tRNA genes were identified by tRNA-SEscan [25] using default parameters. A total of 592 tRNA genes decoding 20 amino acids were detected, together with 15 predicted pseudo- tRNA genes and 7 tRNA genes with an unknown isotype.

Supplementary Table 6. Comparison of gene numbers and features of three grass genomes and the dicot Arabidopsis. Gene and exon statistics are shown for gene complements of rice (IRGSP version RAP2), *Brachypodium* (version 1.0) sorghum (version 1.4) and Arabidopsis (TAIR8).

Feature	**Rice**	***Brachypodium***	**Sorghum**	**Arabidopsis (TAIR8)**
	(RAP2)	**(v1.0)**	**(v1.4)**	
Genome assembly size (bp)	382,150,945	271,923,306	738,540,932	119,186,497
Assembled chromosomes (bp)	382,150,945	271,148,425	659,229,367	119,186,497
Unanchored Sequence (bp)	---	774,881	79,311,565	---
Protein coding loci	28,236	25,532 [1]	27,640 [1,2]	26,990[1]
Exons	134,812	140,142	136,658	142,267
Mean exons per gene	4.77	5.49	4.94	5.27
Mean exon size [bp]	364	268	297	280
Median exon size [bp]	165	140	154	155
Mean intron size [bp]	440	391	444	163
Median intron size [bp]	161	146	147	99
Mean gene size with UTR [bp]	3,403	3,336	3,218	2,174
Median gene size with UTR[bp]	2,807	2,643	2,448	1,889
Mean gene size without UTR[bp]	2,467	2,956	2,927	1,857
Median gene size without UTR[bp]	1,812	2,233	2,154	1,553
Mean intergenic region [bp]	10,339	7,311	17,002[2]	2,266
Median intergenic region [bp]	4,349	3,310	4,238[2]	928
Mean Locus density per 100 kb	7.39	9.39	3.74	22.64

[1] For loci comprising predicted alternative splice variants, the longest representative has been selected.

[2] Only *bona fide* gene models of sorghum were considered for this table [20].

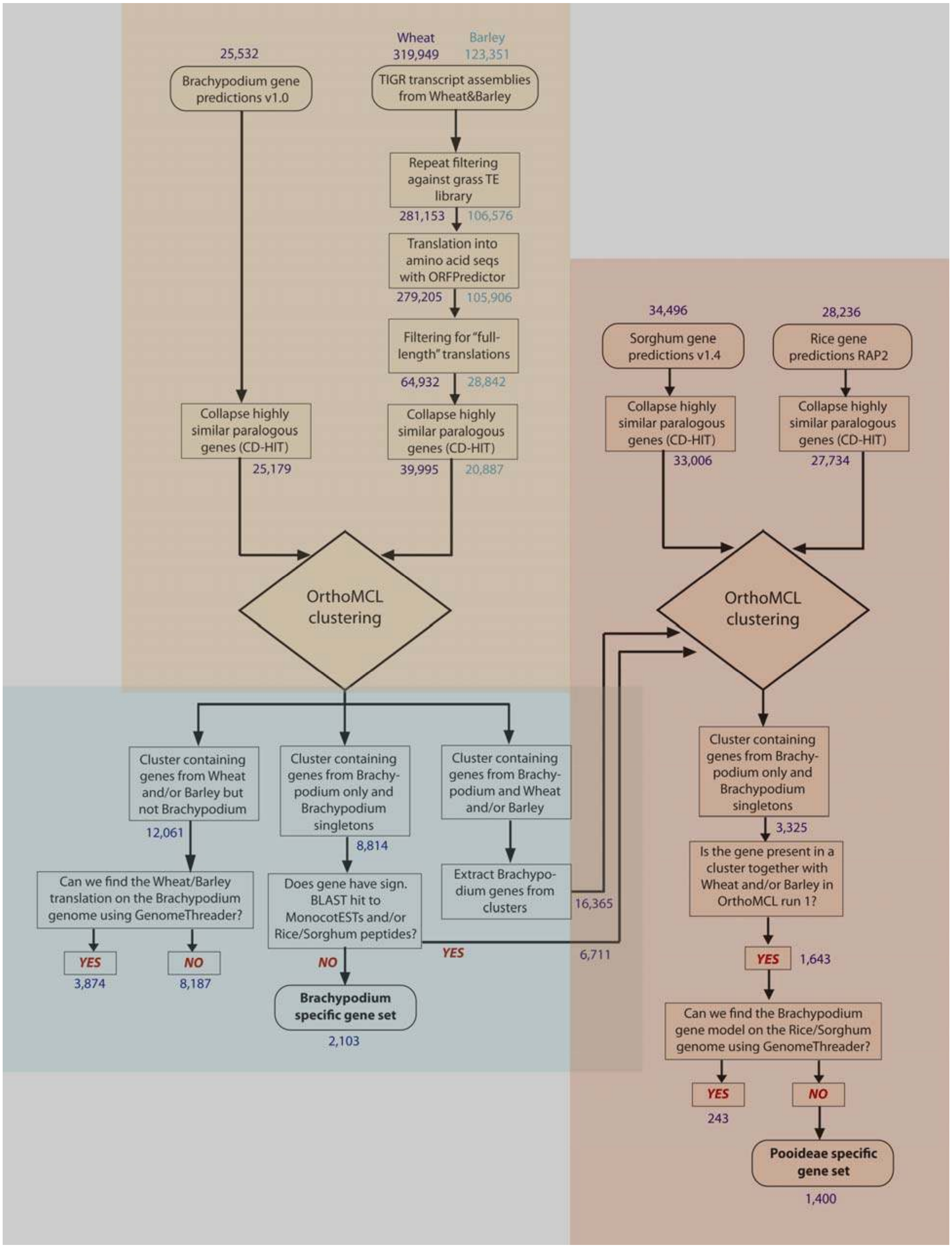


Supplementary Figure 5. Class distribution and extrinsic evidence for *Brachypodium* gene predictions. Initial *Brachypodium* gene predictions were evaluated against supporting evidence from extrinsic data. Gene models were compared against *Brachypodium* ESTs (BdEST), all monocot ESTs from public databases (excluding *Brachypodium*) and Illumina *Brachypodium* transcriptome sequences (Illumina) as well as combinations of these datasets. The fraction of genes in the respective classes (5 highest quality to 0 lowest quality) with supporting extrinsic evidence from the respective resources is depicted in red. Initial gene calls from the classes 0 and 1 without at least 20% overlapping support from extrinsic evidence were filtered from the final v1.0 gene set.

Identification of grass subfamily-specific gene sets

To identify genes and gene families that are enriched in *Brachypodium* and the Pooideae, Ehrhartoideae and Panicoideae subfamilies of the Poaceae we used the *Brachypodium* genome v1.0 gene predictions and multiple EST collections from wheat and barley, as representatives of the Pooideae, the sorghum genome as a representative of the Panicoideae and the rice genome as a representative of the Ehrhartoideae. We applied a rigorous two-way-OrthoMCL clustering scheme along with a data preprocessing to collapse highly similar paralogous genes in the different collections. A flowchart of the data handling steps is given in Supplementary Figure 6. Comparison between *Brachypodium* and wheat and barley transcriptomes was carried out using preprocessed wheat and barley TC/EST dataset that had been repeat filtered, protein translated and filtered for complete reading frame representation. For both *Brachypodium* and the Triticeae dataset highly similar paralogous genes were collapsed using CD-HIT [26]. Due to partial representation, 3,874 wheat/barleyTCs/EST were not grouped with *Brachypodium* genes, although a *Brachypodium* homolog was present. 16,365 *Brachypodium* genes clustered with representatives from wheat /barley and an additional 6,711 had homology to additional monocot EST datasets and/or proteins from rice and sorghum. 2,103 *Brachypodium* genes remained. EST and Illumina sequence data demonstrated that over 80% of these genes were transcribed.

The combined datasets of *Brachypodium*, wheat and barley were clustered against rice and sorghum datasets that were pre-processed to collapse expanded paralogous gene families. 13,580 gene families containing representatives from all three lineages were detected. 681 families were shared between *Brachypodium* and rice (Ehrhartoideae*)* but not with sorghum, and 1,689 families were shared between *Brachypodium* and sorghum but not with rice. 265 families containing 811 genes (1,643 including singleton genes) appeared to have homologs in wheat and barley but not in rice or sorghum and were a potential set of Pooideae- specific genes. However comparison against the rice and sorghum genomes detected 243 genes among them that had homologous loci in rice and/or sorghum that had not been identified previously. This further reduced the number of Pooideae- specific genes without counterparts in rice and sorghum to 1,400 (5.6%). A Venn diagram representing this data is shown in Figure 2C.


Wheat
319,949
Barley
123,351
25,532
Brachypodium gene predictions v1.0
TIGR transcript assemblies from Wheat&Barley
Repeat filtering against grass TE library
281,153
106,576
Translation into amino acid seqs with ORFPredictor
279,205
105,906
Filtering for "full-length" translations
64,932
28,842
Collapse highly similar paralogous genes (CD-HIT)
25,179
Collapse highly similar paralogous genes (CD-HIT)
39,995
20,887
OrthoMCL clustering
34,496
Sorghum gene predictions v1.4
28,236
Rice gene predictions RAP2
Collapse highly similar paralogous genes (CD-HIT)
33,006
Collapse highly similar paralogous genes (CD-HIT)
27,734
OrthoMCL clustering
Cluster containing genes from Wheat and/or Barley but not Brachypodium
12,061
Cluster containing genes from Brachypodium only and Brachypodium singletons
8,814
Cluster containing genes from Brachypodium and Wheat and/or Barley
Extract Brachypodium genes from clusters
16,365
Can we find the Wheat/Barley translation on the Brachypodium genome using GenomeThreader?
YES
3,874
NO
8,187
Does gene have sign. BLAST hit to MonocotESTs and/or Rice/Sorghum peptides?
YES
6,711
NO
Brachypodium specific gene set
2,103
Cluster containing genes from Brachypodium only and Brachypodium singletons
3,325
Is the gene present in a cluster together with Wheat and/or Barley in OrthoMCL run 1?
YES
1,643
Can we find the Brachypodium gene model on the Rice/Sorghum genome using GenomeThreader?
YES
243
NO
Pooideae specific gene set
1,400

Supplementary Figure 6. Workflow of two-way orthoMCL analysis to detect *Brachypodium*- and Pooideae-specific genes.

Grass family and species- specific gene functional categories

The blast2go suite[27] was used to assign molecular functions to gene predictions. 16,589 loci were associated with at least one GO term and a total of 9,086 distinct GO identifiers were mapped onto the v1.0 gene set. The significance of overrepresented GO terms in gene groups was evaluated using the hypergeometric test as implemented in R, and p-values were Bonferroni-corrected for multiple hypothesis testing. We report only results for which at least 20 distinct loci in the full set and at least 5 distinct genes in the relation data set were associated with the respective GO term. In all cases, relations were contrasted to all *Brachypodium* genes that participated in the respective experiment and were associated with GO terms. Enrichment analysis was carried out for specific gene groups of interest obtained from the OrthoMCL analysis described in Supplementary Figure 6, and for tandem repeat genes described in Supplementary Figure 7 below.

Supplementary Table 7. Gene function enrichment in the grasses. Functional categories, indicated by their unique GO identifier in the first column and a short description in the last column, are sorted by decreasing significance (column 4). Related or correlated functional categories are highlighted with the same background colour, which are specific for each table. The second column lists the number of all *Brachypodium* protein coding loci that were included in the respective experiment and that share the category of the first column. The third column shows how many of these genes were observed in the selected group. Results for different selected gene sets are shown.

A. Four-species comparisons that harbour orthologs in Arabidopsis, *Brachypodium*, sorghum and rice, describing angiosperm-specific gene functional categories.
B. Grass-specific orthologs that are shared in *Brachypodium*, sorghum and rice but lack a detectable ortholog in Arabidopsis.
C. A set of Pooideae- specific orthologs that were obtained by the OrthoMCL scheme described in Supplementary Figure 6.
D. *Brachypodium* specific gene functional categories.

7A. Angiosperm-specific gene functions

GO-ID	#genes in Bd	#genes in group	pvalue	GO description
GO:0005515	9363	6528	3.732445e-037	protein binding
GO:0017111	1358	1092	7.540423e-033	nucleoside-triphosphatase activity
GO:0016462	1424	1136	1.815919e-031	pyrophosphatase activity
GO:0016818	1431	1140	4.201848e-031	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides
GO:0016817	1440	1143	6.293848e-030	hydrolase activity, acting on acid anhydrides
GO:0016887	1041	844	6.925815e-027	ATPase activity
GO:0042623	826	683	5.312438e-026	ATPase activity, coupled
GO:0015405	255	233	3.291337e-019	P-P-bond-hydrolysis-driven transmembrane transporter activity
GO:0003723	1155	903	4.673948e-019	RNA binding
GO:0015399	263	238	3.404719e-018	primary active transmembrane transporter activity
GO:0043492	230	211	8.486391e-018	ATPase activity, coupled to movement of substances
GO:0042626	221	203	3.115474e-017	ATPase activity, coupled to transmembrane movement of substances
GO:0022892	1331	1017	5.518591e-016	substrate-specific transporter activity
GO:0005215	1527	1153	1.824022e-015	transporter activity
GO:0016787	3652	2613	4.158226e-015	hydrolase activity
GO:0003735	297	259	1.227306e-014	structural constituent of ribosome
GO:0022804	784	620	2.509192e-014	active transmembrane transporter activity
GO:0016820	229	205	4.043024e-014	hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances
GO:0022857	1233	940	4.279069e-014	transmembrane transporter activity
GO:0022891	1089	828	1.197746e-011	substrate-specific transmembrane transporter activity
GO:0005198	775	603	3.087713e-011	structural molecule activity
GO:0000166	3223	2293	8.237713e-011	nucleotide binding
GO:0015075	810	626	1.102622e-010	ion transmembrane transporter activity
GO:0008324	678	529	5.154825e-010	cation transmembrane transporter activity
GO:0017076	2815	2006	2.133855e-009	purine nucleotide binding
GO:0022890	352	289	3.350292e-009	inorganic cation transmembrane transporter activity
GO:0003824	9280	6294	3.820845e-009	catalytic activity
GO:0032555	2661	1886	2.401375e-007	purine ribonucleotide binding
GO:0032553	2661	1886	2.401375e-007	ribonucleotide binding
GO:0008028	90	84	4.138798e-007	monocarboxylic acid transmembrane transporter activity
GO:0051082	253	210	4.285327e-007	unfolded protein binding
GO:0042625	125	112	4.761162e-007	ATPase activity, coupled to transmembrane movement of ions
GO:0005319	139	123	4.776445e-007	lipid transporter activity
GO:0050662	407	323	5.644548e-007	coenzyme binding
GO:0015239	71	68	7.500193e-007	multidrug transporter activity
GO:0015662	112	101	1.405945e-006	ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism
GO:0001882	2640	1863	2.882158e-006	nucleoside binding
GO:0015238	180	153	3.330429e-006	drug transporter activity
GO:0001883	2630	1854	5.293324e-006	purine nucleoside binding
GO:0030554	2602	1832	1.200137e-005	adenyl nucleotide binding
GO:0046873	346	274	1.629670e-005	metal ion transmembrane transporter activity
GO:0008017	264	214	1.728550e-005	microtubule binding
GO:0048037	539	412	2.146384e-005	cofactor binding
GO:0008135	159	135	3.149459e-005	translation factor activity, nucleic acid binding
GO:0045182	199	165	3.522271e-005	translation regulator activity
GO:0008565	182	152	4.569390e-005	protein transporter activity
GO:0004386	240	195	5.145681e-005	helicase activity
GO:0043021	156	132	6.895191e-005	ribonucleoprotein binding
GO:0016853	291	231	1.443650e-004	isomerase activity
GO:0015631	405	312	2.887190e-004	tubulin binding
GO:0005548	84	75	4.922353e-004	phospholipid transporter activity
GO:0043022	74	67	5.707748e-004	ribosome binding
GO:0008026	194	158	6.742438e-004	ATP-dependent helicase activity
GO:0070035	194	158	6.742438e-004	purine NTP-dependent helicase activity
GO:0015082	151	126	6.816331e-004	di-, tri-valent inorganic cation transmembrane transporter activity
GO:0016810	151	126	6.816331e-004	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds
GO:0019829	78	70	7.390135e-004	cation-transporting ATPase activity
GO:0032559	2449	1712	7.721869e-004	adenyl ribonucleotide binding
GO:0051536	100	87	9.100646e-004	iron-sulfur cluster binding
GO:0051540	100	87	9.100646e-004	metal cluster binding
GO:0003743	72	65	1.092129e-003	translation initiation factor activity
GO:0005525	262	207	1.171458e-003	GTP binding
GO:0016638	39	38	1.277029e-003	oxidoreductase activity, acting on the CH-NH2 group of donors
GO:0015432	38	37	1.897058e-003	bile acid-exporting ATPase activity
GO:0034040	38	37	1.897058e-003	lipid-transporting ATPase activity
GO:0050660	137	114	2.978936e-003	FAD binding
GO:0046915	137	114	2.978936e-003	transition metal ion transmembrane transporter activity
GO:0005342	267	209	3.496742e-003	organic acid transmembrane transporter activity
GO:0045502	116	98	3.676276e-003	dynein binding
GO:0005083	218	173	5.070466e-003	small GTPase regulator activity
GO:0046943	254	199	5.282991e-003	carboxylic acid transmembrane transporter activity
GO:0015125	57	52	6.085054e-003	bile acid transmembrane transporter activity
GO:0008649	28	28	6.087728e-003	rRNA methyltransferase activity
GO:0016407	176	142	6.518986e-003	acetyltransferase activity
GO:0008144	84	73	7.328962e-003	drug binding
GO:0042803	595	439	8.211977e-003	protein homodimerization activity
GO:0008173	56	51	8.487498e-003	RNA methyltransferase activity
GO:0032561	297	229	9.040328e-003	guanyl ribonucleotide binding
GO:0016410	136	112	9.676780e-003	N-acyltransferase activity
GO:0008415	317	243	1.053337e-002	acyltransferase activity
GO:0003924	162	131	1.163203e-002	GTPase activity
GO:0046527	95	81	1.194623e-002	glucosyltransferase activity
GO:0008757	206	163	1.270669e-002	S-adenosylmethionine-dependent methyltransferase activity
GO:0016741	326	249	1.317692e-002	transferase activity, transferring one-carbon groups
GO:0019001	298	229	1.370324e-002	guanyl nucleotide binding
GO:0015077	183	146	1.552005e-002	monovalent inorganic cation transmembrane transporter activity
GO:0035254	44	41	1.586072e-002	glutamate receptor binding
GO:0016866	54	49	1.643252e-002	intramolecular transferase activity
GO:0004004	89	76	1.951860e-002	ATP-dependent RNA helicase activity
GO:0008186	97	82	2.095847e-002	RNA-dependent ATPase activity
GO:0034634	25	25	2.147404e-002	glutathione transmembrane transporter activity
GO:0015248	48	44	2.351942e-002	sterol transporter activity
GO:0005524	2293	1591	2.564999e-002	ATP binding
GO:0003774	287	220	2.658952e-002	motor activity
GO:0035251	75	65	2.777867e-002	UDP-glucosyltransferase activity
GO:0008168	321	244	2.886193e-002	methyltransferase activity
GO:0008553	42	39	3.210211e-002	hydrogen-exporting ATPase activity, phosphorylative mechanism
GO:0004705	24	24	3.268658e-002	JUN kinase activity
GO:0016251	70	61	3.359403e-002	general RNA polymerase II transcription factor activity
GO:0004437	65	57	4.010906e-002	inositol or phosphatidylinositol phosphatase activity
GO:0016814	30	29	4.379283e-002	hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in cyclic amidines
GO:0042277	210	164	4.571028e-002	peptide binding
GO:0030695	338	255	4.880359e-002	GTPase regulator activity
GO:0016908	23	23	4.975192e-002	MAP kinase 2 activity

7B. Grass-specific gene functions

GO-ID	#loci in Bd	#loci in group	pvalue	GO description
GO:0019199	296	118	1.090913e-012	transmembrane receptor protein kinase activity
GO:0005149	517	178	2.410879e-012	interleukin-1 receptor binding
GO:0004714	175	79	2.063034e-011	transmembrane receptor protein tyrosine kinase activity
GO:0015020	114	59	2.172584e-011	glucuronosyltransferase activity
GO:0008083	545	182	3.060555e-011	growth factor activity
GO:0046906	383	135	7.726515e-010	tetrapyrrole binding
GO:0020037	378	133	1.314941e-009	heme binding
GO:0005003	79	42	3.259965e-008	ephrin receptor activity
GO:0016757	557	172	2.099423e-007	transferase activity, transferring glycosyl groups
GO:0046914	2116	529	4.876136e-007	transition metal ion binding
GO:0043167	3445	813	1.089482e-006	ion binding
GO:0043169	3426	808	1.436583e-006	cation binding
GO:0016563	1152	309	2.060071e-006	transcription activator activity
GO:0016758	435	137	3.374423e-006	transferase activity, transferring hexosyl groups
GO:0019904	969	264	6.790280e-006	protein domain specific binding
GO:0004888	548	163	1.145715e-005	transmembrane receptor activity
GO:0046872	3284	768	2.200052e-005	metal ion binding
GO:0005057	646	185	2.809617e-005	receptor signaling protein activity
GO:0005506	537	158	4.146637e-005	iron ion binding
GO:0004872	678	191	6.288092e-005	receptor activity
GO:0004713	1012	267	1.252530e-004	protein tyrosine kinase activity
GO:0008194	323	103	1.310114e-004	UDP-glycosyltransferase activity
GO:0016684	173	63	1.941035e-004	oxidoreductase activity, acting on peroxide as acceptor
GO:0004601	173	63	1.941035e-004	peroxidase activity
GO:0004702	549	157	3.214428e-004	receptor signaling protein serine/threonine kinase activity
GO:0004709	312	98	5.507980e-004	MAP kinase kinase kinase activity
GO:0003700	768	205	1.463517e-003	transcription factor activity
GO:0043565	655	177	3.230145e-003	sequence-specific DNA binding
GO:0016209	240	77	3.255155e-003	antioxidant activity
GO:0008395	175	59	7.417863e-003	steroid hydroxylase activity
GO:0004497	293	89	7.503592e-003	monooxygenase activity
GO:0016505	49	23	1.070010e-002	apoptotic protease activator activity
GO:0005102	1420	344	1.402805e-002	receptor binding
GO:0016504	53	24	1.499831e-002	peptidase activator activity
GO:0003704	119	43	1.651342e-002	specific RNA polymerase II transcription factor activity
GO:0009055	668	175	2.460068e-002	electron carrier activity
GO:0046332	155	52	2.718835e-002	SMAD binding
GO:0008301	56	24	4.479893e-002	DNA bending activity
GO:0035250	56	24	4.479893e-002	UDP-galactosyltransferase activity

7C. Pooid- specific gene functions

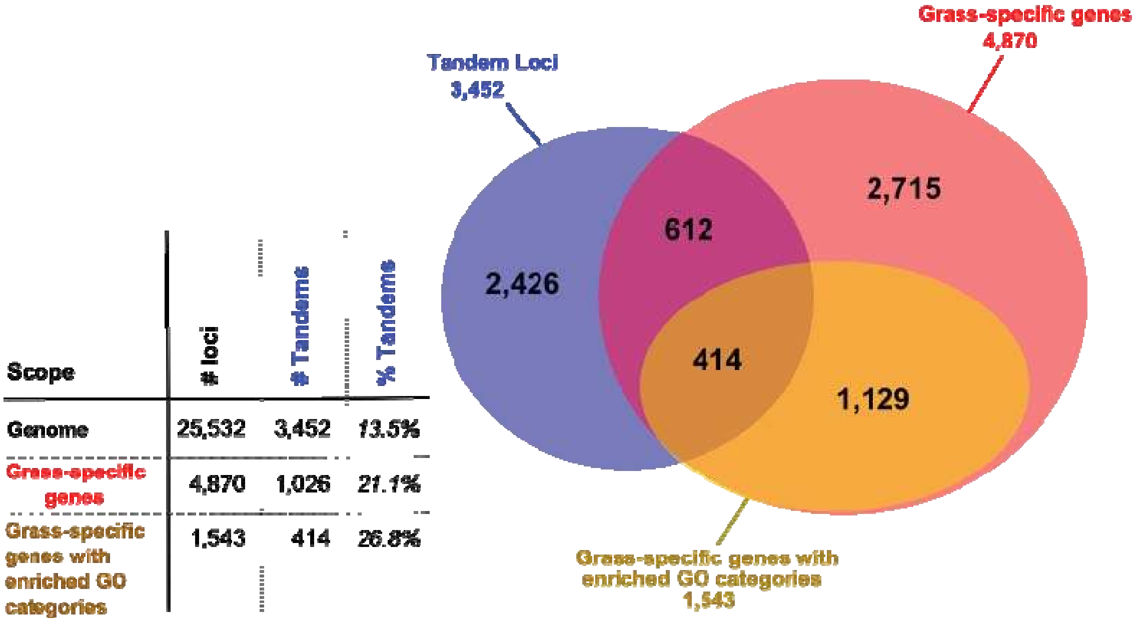
GO-ID	#genes in Bd	#genes in group	pvalue	GO description
GO:0016684	173	24	1.117948e-007	oxidoreductase activity, acting on peroxide as acceptor
GO:0004601	173	24	1.117948e-007	peroxidase activity
GO:0016209	240	24	6.846456e-005	antioxidant activity
GO:0004867	26	8	1.149002e-004	serine-type endopeptidase inhibitor activity
GO:0020037	378	30	3.704022e-004	heme binding
GO:0046906	383	30	4.835093e-004	tetrapyrrole binding
GO:0004185	56	10	1.103453e-003	serine-type carboxypeptidase activity
GO:0070008	56	10	1.103453e-003	serine-type exopeptidase activity
GO:0046914	2116	98	3.075212e-003	transition metal ion binding
GO:0004180	70	10	8.345396e-003	carboxypeptidase activity
GO:0008233	686	40	1.401720e-002	peptidase activity
GO:0004866	90	11	1.546043e-002	endopeptidase inhibitor activity
GO:0030414	93	11	2.084435e-002	peptidase inhibitor activity
GO:0005506	537	33	2.222067e-002	iron ion binding

7D. *Brachypodium*-specific gene functions

GO-ID	#genes in Bd	#genes in group	pvalue	GO description
GO:0016684	173	24	1.117948e-007	oxidoreductase activity, acting on peroxide as acceptor
GO:0004601	173	24	1.117948e-007	peroxidase activity
GO:0016209	240	24	6.846456e-005	antioxidant activity
GO:0004867	26	8	1.149002e-004	serine-type endopeptidase inhibitor activity
GO:0020037	378	30	3.704022e-004	heme binding
GO:0046906	383	30	4.835093e-004	tetrapyrrole binding
GO:0004185	56	10	1.103453e-003	serine-type carboxypeptidase activity
GO:0070008	56	10	1.103453e-003	serine-type exopeptidase activity
GO:0046914	2116	98	3.075212e-003	transition metal ion binding
GO:0004180	70	10	8.345396e-003	carboxypeptidase activity
GO:0008233	686	40	1.401720e-002	peptidase activity
GO:0004866	90	11	1.546043e-002	endopeptidase inhibitor activity
GO:0030414	93	11	2.084435e-002	peptidase inhibitor activity
GO:0005506	537	33	2.222067e-002	iron ion binding

Identification of tandem repeat genes

An undirected graph with genes as nodes and protein similarities as edge weights was constructed for the *Brachypodium* protein coding gene set v1.0. Protein similarities were derived from pair-wise local Smith-Waterman alignments (blastp). An e-value $\leq 10^{-15}$ and a minimal alignment coverage of ≥70% of both protein sizes were required. Edges connecting genes that were more than 9 genes distant from each other in the genome were removed and tandem clusters were retrieved as connected groups from the resulting graph. In total, we detected 1,313 clusters comprising 3,452 (13.5% of all *Brachypodium* genes) tandem repeated genes. The gene classes enriched in pooid- and *Brachypodium*- core sets had a highly significant increased proportion of tandem genes, 21.1% compared to 13.5% in the whole genome.



Scope	# loci	# Tandems	% Tandems
Genome	25,532	3,452	*13.5%*
Grass-specific genes	4,870	1,026	*21.1%*
Grass-specific genes with enriched GO categories	1,543	414	*26.8%*

Supplementary Figure 7. Tandemly repeated genes contribute disproportionately to grass- specific gene functions in *Brachypodium*. Tandem genes (blue circle) comprise 3,452 loci (13.5%) out of 25,532 loci (see Supplementary Figure 6). This proportion was used to test the hypothesis that genes categorized as grass-specific genes were enriched for tandem duplications in *Brachypodium*. The significance was tested by one-sided Fisher's exact test as implemented in R (http://www.r-project.org/). 4,870 *Brachypodium* loci (red circle) were detected in the four-way OrthoMCL analysis as grass-specific genes. 1,026 (21.1%) of these are tandemly duplicated genes, as shown by the intersection of the red and the blue circles. The increased representation of tandem genes in grass-specific genes is highly significant ($p<10^{-16}$). The increased proportion of tandem genes was even more pronounced for those grass-specific genes that were associated with significantly enriched GO functional categories. Out of 4,870 grass- specific genes, 1,543 were associated with enriched categories (light brown circle, strict subset of grass core). 414 (26.8%) of these genes were tandemly repeated genes suggesting that tandem duplication is an important mechanism for generating grass- specific gene functions.

Supplementary Table 8. Gene functions enriched in tandemly repeated genes. Functional categories enriched in tandem genes are shown grouped by GO identifiers (column 1). The second column lists the number of genes in the *Brachypodium* genome annotated with the GO id and the third column lists the number of tandemly repeated genes with the GO id. Enriched categories are sorted by decreasing significance (4th column). Background colours highlight related GO terms that are either parent-child relations or have widely overlapping functions.

GO_ID	#genes in Bd	#genes in group	pvalue	GO description
GO:0005149	579	258	2.139709e-053	interleukin-1 receptor binding
GO:0008083	613	262	6.093913e-050	growth factor activity
GO:0004888	623	263	6.927789e-049	transmembrane receptor activity
GO:0004713	1114	380	3.215135e-044	protein tyrosine kinase activity
GO:0004872	763	292	3.845380e-044	receptor activity
GO:0020037	473	211	3.276159e-043	heme binding
GO:0046906	479	212	9.310332e-043	tetrapyrrole binding
GO:0019199	343	166	3.775488e-039	transmembrane receptor protein kinase activity
GO:0009055	793	289	7.901405e-039	electron carrier activity
GO:0004714	212	121	2.317908e-037	transmembrane receptor protein tyrosine kinase activity
GO:0005506	645	242	2.175601e-034	iron ion binding
GO:0004674	1356	402	8.408213e-031	protein serine/threonine kinase activity
GO:0004871	1601	453	6.849715e-030	signal transducer activity
GO:0060089	1601	453	6.849715e-030	molecular transducer activity
GO:0004672	1524	427	6.810287e-027	protein kinase activity
GO:0016491	1712	454	4.641683e-023	oxidoreductase activity
GO:0016684	206	100	4.961243e-023	oxidoreductase activity, acting on peroxide as acceptor
GO:0004601	206	100	4.961243e-023	peroxidase activity
GO:0005102	1591	428	6.925593e-023	receptor binding
GO:0005057	714	233	9.192004e-023	receptor signaling protein activity
GO:0004702	605	204	1.233903e-021	receptor signaling protein serine/threonine kinase activity
GO:0004497	364	142	3.045009e-021	monooxygenase activity
GO:0008395	218	100	1.193911e-020	steroid hydroxylase activity
GO:0016209	279	117	2.704615e-020	antioxidant activity
GO:0016773	1701	435	1.479608e-018	phosphotransferase activity, alcohol group as acceptor
GO:0005003	88	54	6.717141e-018	ephrin receptor activity
GO:0019904	1063	292	4.589120e-016	protein domain specific binding
GO:0008391	146	70	2.515367e-015	arachidonic acid monooxygenase activity
GO:0016705	392	136	5.042972e-015	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
GO:0016301	1798	439	9.274991e-015	kinase activity
GO:0004709	340	122	1.302847e-014	MAP kinase kinase kinase activity
GO:0005524	2512	577	2.813425e-014	ATP binding
GO:0045735	66	40	1.232182e-012	nutrient reservoir activity
GO:0043169	3784	807	1.384161e-012	cation binding
GO:0043167	3806	808	4.315672e-012	ion binding
GO:0032559	2675	590	6.485347e-011	adenyl ribonucleotide binding
GO:0005529	372	121	6.883335e-011	sugar binding
GO:0046872	3633	766	1.702493e-010	metal ion binding
GO:0003824	10325	1925	2.034155e-010	catalytic activity
GO:0046914	2376	527	7.712537e-010	transition metal ion binding
GO:0015020	125	55	7.981506e-010	glucuronosyltransferase activity
GO:0030246	488	145	8.145736e-010	carbohydrate binding
GO:0030554	2845	614	1.386625e-009	adenyl nucleotide binding
GO:0015197	94	45	2.408946e-009	peptide transporter activity
GO:0019865	70	37	4.517551e-009	immunoglobulin binding
GO:0001883	2877	616	5.868084e-009	purine nucleoside binding
GO:0016758	497	144	7.726534e-009	transferase activity, transferring hexosyl groups
GO:0001882	2887	616	1.135329e-008	nucleoside binding
GO:0015198	85	40	7.205140e-008	oligopeptide transporter activity
GO:0016772	2051	453	8.454987e-008	transferase activity, transferring phosphorus-containing groups
GO:0005178	127	52	8.474492e-008	integrin binding
GO:0019863	55	30	1.536028e-007	IgE binding
GO:0016740	3808	777	1.616598e-007	transferase activity
GO:0004568	36	23	2.524012e-007	chitinase activity
GO:0000016	29	20	4.616243e-007	lactase activity
GO:0032403	505	139	8.147338e-007	protein complex binding
GO:0032555	2905	602	3.399175e-006	purine ribonucleotide binding
GO:0032553	2905	602	3.399175e-006	ribonucleotide binding
GO:0031013	190	65	3.510434e-006	troponin I binding
GO:0004706	112	44	9.665124e-006	JUN kinase kinase kinase activity
GO:0050839	63	30	1.037699e-005	cell adhesion molecule binding
GO:0008422	30	19	1.053577e-005	beta-glucosidase activity
GO:0005507	160	56	1.564732e-005	copper ion binding
GO:0016757	622	158	2.675785e-005	transferase activity, transferring glycosyl groups
GO:0050649	40	22	3.059046e-005	testosterone 6-beta-hydroxylase activity
GO:0017076	3077	626	3.094471e-005	purine nucleotide binding
GO:0030304	25	16	1.167670e-004	trypsin inhibitor activity
GO:0016563	1256	281	1.306390e-004	transcription activator activity
GO:0004866	110	41	1.606202e-004	endopeptidase inhibitor activity
GO:0030414	113	41	3.697257e-004	peptidase inhibitor activity
GO:0004033	80	32	5.204065e-004	aldo-keto reductase activity
GO:0008194	346	94	6.909026e-004	UDP-glycosyltransferase activity
GO:0004185	70	29	7.092495e-004	serine-type carboxypeptidase activity
GO:0070008	70	29	7.092495e-004	serine-type exopeptidase activity
GO:0004704	78	31	9.032958e-004	NF-kappaB-inducing kinase activity
GO:0004553	367	98	9.552263e-004	hydrolase activity, hydrolyzing O-glycosyl compounds
GO:0015238	189	58	1.378798e-003	drug transporter activity
GO:0016682	26	15	1.860800e-003	oxidoreductase activity, acting on diphenols and related substances as donors, oxygen as acceptor
GO:0004180	85	32	2.458195e-003	carboxypeptidase activity
GO:0045295	39	19	2.778457e-003	gamma-catenin binding
GO:0008390	24	14	3.336766e-003	testosterone 16-alpha-hydroxylase activity
GO:0004032	27	15	3.527146e-003	aldehyde reductase activity
GO:0004869	76	29	5.018050e-003	cysteine-type endopeptidase inhibitor activity
GO:0005427	34	17	5.634335e-003	proton-dependent oligopeptide secondary active transmembrane transporter activity
GO:0015322	34	17	5.634335e-003	secondary active oligopeptide transmembrane transporter activity
GO:0008378	92	33	5.860704e-003	galactosyltransferase activity
GO:0008061	22	13	5.973128e-003	chitin binding
GO:0035250	62	25	6.694382e-003	UDP-galactosyltransferase activity
GO:0004508	45	20	8.979709e-003	steroid 17-alpha-monooxygenase activity
GO:0005504	98	34	9.839197e-003	fatty acid binding
GO:0000287	688	159	1.020294e-002	magnesium ion binding
GO:0042895	20	12	1.066603e-002	antibiotic transporter activity
GO:0016762	23	13	1.159414e-002	xyloglucan:xyloglucosyl transferase activity
GO:0030145	166	50	1.162094e-002	manganese ion binding
GO:0008545	26	14	1.168276e-002	JUN kinase kinase activity
GO:0019838	80	29	1.571803e-002	growth factor binding
GO:0045296	43	19	1.614062e-002	cadherin binding
GO:0015239	73	27	1.947246e-002	multidrug transporter activity
GO:0015293	215	60	2.385704e-002	symporter activity
GO:0016709	70	26	2.485116e-002	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, NADH
GO:0033293	123	39	2.611737e-002	monocarboxylic acid binding
GO:0004708	94	32	2.635998e-002	MAP kinase kinase activity
GO:0015925	48	20	2.818589e-002	galactosidase activity
GO:0004565	45	19	3.479884e-002	beta-galactosidase activity

Manual annotation and gene family analysis

Gene models (2,755) from gene families or pathways were selected for manual annotation based on BLAST scores to known genes and/or from the presence of pfam domains (Supplementary Table 9). We emphasized gene families relevant to bioenergy research, including genes involved in the biosynthesis and remodeling of the cell wall (cellulose synthase (10 genes), cellulose synthase-like (CSL, 25 genes), other glycosyltransferases (313 genes), glycosyl hydrolases (339 genes), and 179 genes putatively involved in monolignol or pectin metabolism. Selected genes were manually examined and edited using EST alignments, Illumina transcriptome data, splice site verification by Illumina sequence and alignment to previously described genes from other organisms. Phylogenetic analysis of 62 gene families demonstrated that in most cases *Brachypodium*, rice and sorghum had very similar gene family compositions. One surprising example involves the CSL sub-family J, which was recently proposed to be present in some grasses including maize, sorghum, barley, and wheat, but not in others including rice and *Brachypodium*, or dicots [28]. Our analysis confirmed the absence of CSLJ genes from *Brachypodium* and rice, although it did reveal the presence of CSLJ in poplar and several other dicots (Supplementary Figure 9).

Glycosyltransferases (GTs) related to cell wall biosynthesis and many other cell functions are generally conserved between angiosperms. 40 GT families have representatives in all angiosperms that have been analyzed to date. In the GT4-GT90 families there are 310 members in *Brachypodium*, 316 in rice and 291 in Arabidopsis. In most cases the phylogenetic trees reveal clear orthologs in all three species, with the occasional duplication of genes in only one of the species. Notable exceptions to this common picture are found in a few GT families, GT37, GT43 and GT61, which have significantly more GTs in the grasses. Interestingly, the opposite is not the case – there are no GT families where Arabidopsis has many more members than the two grasses. At the subfamily level we found some clades with no or very few grass members, e.g. in GT37. GT37 includes xyloglucan fucosyltransferases, but it appears that only one of the 10 Arabidopsis genes encodes an enzyme with this activity, while some or all of the rest encode other fucosyltransferases [29]. *Brachypodium* and rice have 16 and 18 members of GT37, respectively, but only one GT from each of these species clusters, with poor resolution, with the 10 Arabidopsis genes, and not as orthologs to the known xyloglucan fucosyltransferase. Fucosylated xyloglucan is usually not found in grasses. The other GT37 members in rice and *Brachypodium* do not cluster into clearly defined subfamilies, but they do form 11-12 separate clades, which all contain both rice and *Brachypodium* orthologs. This means that whatever function these grass-specific GT37 members have, they had evolved their special functions in the common ancestor of rice and *Brachypodium*.

GT43 is known to contain GTs involved in xylan biosynthesis. In Arabidopsis the 4 genes fall in two groups; the ‘A-group’ containing *irx9* and a homolog of *irx9* and a ‘B-group’ containing *irx14* and a homolog of *irx14*. These seem to be involved in xylan biosynthesis, and genetic evidence suggests that xylan synthase may require a member from each group. Rice and *Brachypodium* both have 10 GT43 members, two in the A-group and 8 in the B-group, although some of these GTs are quite diverged from Arabidopsis *irx14*. All the GT43 members in rice and *Brachypodium* occur in clearly orthologous pairs, indicating that as for GT37, specialization had occurred already in the common ancestor. The functions of the grass specific GT43 members are not yet known, but it seems reasonable to assume that they are all involved in xylan biosynthesis and that the grass specific groups have functions related to the important role of xylans in grass primary walls.

GT61 contains a protein *N*-glycan xylosyltransferase gene, which falls in a distinct B-clade with one member in Arabidopsis, rice and *Brachypodium* (BdXYLT). The other GT61 members fall in a well-defined ‘A-group’ and a diverse ‘C-group’, which cannot easily be divided into well-defined clades. Arabidopsis, rice and *Brachypodium* all have 4 members of GT61A, but the grass members are not apparent orthologs of the Arabidopsis members. However, the rice and *Brachypodium* members of GT61A form four orthologous pairs. The C-group is very abundant in grasses with 16 and 20 members in *Brachypodium* and rice, respectively, and only two in Arabidopsis. There are clearly orthologous pairs for most of the rice and *Brachypodium* GT family members, indicating an early diversification. Some of the GT61 members in rice are known to be coexpressed with xylan biosynthetic genes [30], and they are therefore good candidates for xylan arabinosyltransferases. It is unknown if Arabidopsis has any arabinose substitutions on xylan, but such substitutions are known from other dicots. Perhaps all the GT61 members – except for the *N*-glycan xylosyltransferase – are involved in arabinosylation of xylan, and the great diversification in grasses signify the different patterns of arabinosylation and the importance of arabinoxylan in grasses.

In conclusion, *Brachypodium* and rice have a very similar set of GTs. There is no evidence for 'rice-specific' or '*Brachypodium*-specific' GTs, but only 'grass specific' GTs. This is consistent with the analyses shown in Supplementary Table 7. At least for GT43 and GT61 the evidence suggests that the grass specific GTs are related to xylan biosynthesis. The overrepresentation makes sense in view of the very important role of xylans as the main matrix polysaccharide in primary walls of grasses – a role which is filled by xyloglucan and pectins in other plants. The evolution of the specialized grass cell wall has led to a diversification of a limited set of GTs, and this appears to have been a key event that took place very early in evolution of grasses. It should be noted that a xylan rich primary wall is found also in some other commelinid species besides grasses, but none of these species have yet been analyzed at the genome sequence level. It is perhaps surprising that although pectins and xyloglucans are present at much lower levels in grass cell walls than in plants like Arabidopsis, the grasses and dicots retain a similar number of genes thought to be involved in their synthesis. With regard to pectins it should be borne in mind that synthesis of the pectic middle lamella is indispensable to cell division also in grasses. It may be hypothesized that the process of de novo wall formation during cell division generally is more conserved among angiosperm families than the mature primary wall structure is, and also that xyloglucan is required for wall assembly during cytokinesis in grasses. In addition, besides the roles of pectins and xyloglucan as 'bulk' matrix polymers, which would seem of little importance in grasses, these polymers also have roles as a source of signal molecules, which could have prevented their disappearance during grass evolution.

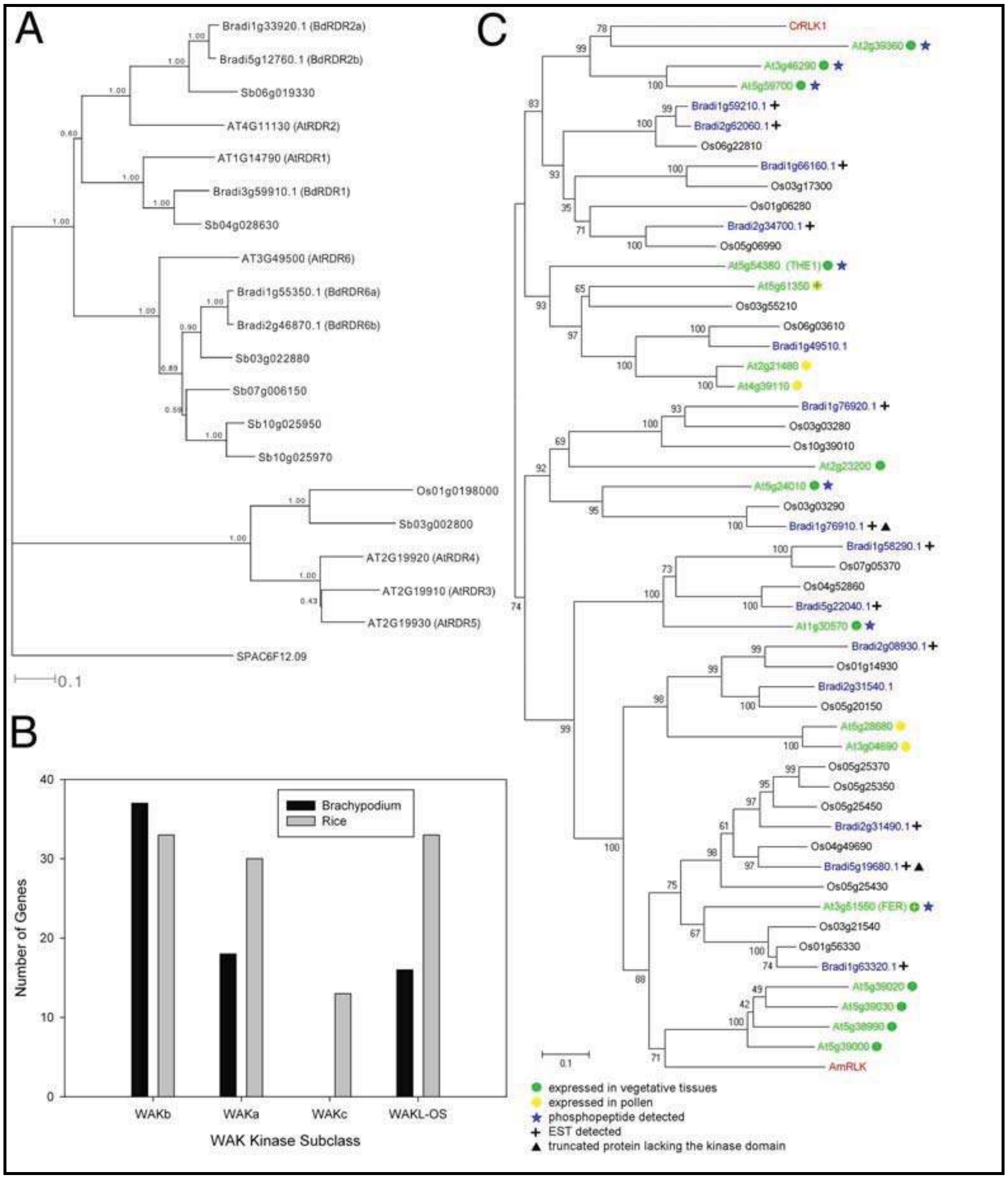
The flowering time pathway is highly conserved and *Brachypodium* contained the expected genes [31] that are also shared by Arabidopsis and rice. However, rice has an additional pathway to effect photoperiodic control of flowering time that utilizes the response regulator Early Heading Date (Ehd) 1 to promote expression of *Hd3* independent of *Hd1*. Day length signals are transmitted by light signaling pathways to control *Ehd1* expression [32]. The Ghd7 transcription factor negatively regulates *Ehd1* expression in response to red light, whereas blue light promotes *Ehd1* expression through the action of the CCT-domain transcription factor Ehd2. Clear orthologs of *Ghd7* and *Ehd2* are present in *Brachypodium*, consistent with some aspects of this flowering pathway being present; however, an obvious *Ehd1* ortholog is missing from the *Brachypodium* genome, despite the identification of *Ehd1* orthologs in sorghum and maize. Thus, the structure of this pathway in *Brachypodium* may be different from rice.

The *RDR* family of genes involved in small RNA processing shows some differences in *Brachypodium*. Rice and sorghum have an ortholog in a clade with the Arabidopsis RDR3, 4, 5 genes while *Brachypodium* does not (Supplementary Figure 8A). Therefore this family member may have been lost in *Brachypodium.* However, *Brachypodium* does have five other RDR genes in the other three RDR clades.

Comparison of the rice, Arabidopsis, and poplar kinomes to the *Brachypodium* kinome (1,177 proteins) demonstrated similar composition to rice (1,454 proteins) but had fewer kinases. Both rice and *Brachypodium* encode the same kinase subfamilies that are very similar in size, with the exception of eight receptor-like kinase (RLK) subfamilies [33]. These subfamilies (LRR-I, DUF26, LRR-VIII-2, LRK10L-2, L-LEC, WAK, LRR-XII, and SD-2b) account for nearly all (252/268) of the total difference in kinome size. The greatest differences were found among the non-RD (arginine-aspartate) kinase subclass that are predicted to encode pattern recognition receptors (LRK10L-2, WAK, LRR-XII, and SD-2b) based on the absence of the conserved R in kinase subdomain VI [34]. These non-RD receptor kinases are under positive selection [34,35]. This is particularly evident among the WAK kinases which contain both RD and non-RD clades. The increased numbers of WAK kinases in rice were almost exclusively among the non-RD WAK class (Supplementary Figure 8B).

The CrRLK1L subfamily of plant-specific proteins (RD kinases) has 17 members in Arabidopsis, 14 in *Brachypodium* and 20 in rice (Supplementary Figure 8C). Seven subclasses were distinguished each with members both in Arabidopsis and rice/*Brachypodium* (except one), indicating that they predate the monocot-dicot split, 160 million years ago. FERONIA is expressed in the synergid cells of the female gametophyte and controls the recognition of the pollen tube [36]. AmRLK is expressed in the petal epidermis of *Antirrhinum* and may be involved in the polar outgrowth of epidermal cells [37]. The FER subclass, which contains a single gene in Arabidopsis, has seven members in rice and three in *Brachypodium*. This could reflect a diversification of pollen tube recognition that may play a role in reproductive isolation within this species. Interestingly, the AmRLK branch contains four tandem-duplicated members in Arabidopsis but none in rice or *Brachypodium* (or in sorghum). This absence may be related to the difference between petals in dicots and lodicules in grasses.

Using BLAST scores and pfam domains, we placed a further 2,749 gene models into 12 gene families including kinases, proteasome subunits, auxin signaling genes and F-box proteins, but these gene models were not manually examined (Supplementary Table 10). Two of these gene families, F-box genes and Bric-a-Brac/Tramtrack/Broad (BTB) Complex, had fewer members than expected based on comparison to other species (Supplementary Table 11). Using domain scans of unmasked genome sequence we identified an additional 62 putative F-box containing genes and 67 putative BTB genes and brought these gene family numbers into a broad agreement with other plants (Supplementary Table 11).



Supplementary Figure 8. Examples of gene families that differed among the grasses. (A) Phylogenetic trees of RDR, (B) Distribution of WAK kinase subfamily members in *Brachypodium* and rice, and (C) Phylogenetic trees of CrRLK1L gene families.

Supplementary Table 9. Manually annotated genes. Genes and gene families that were annotated by experts.

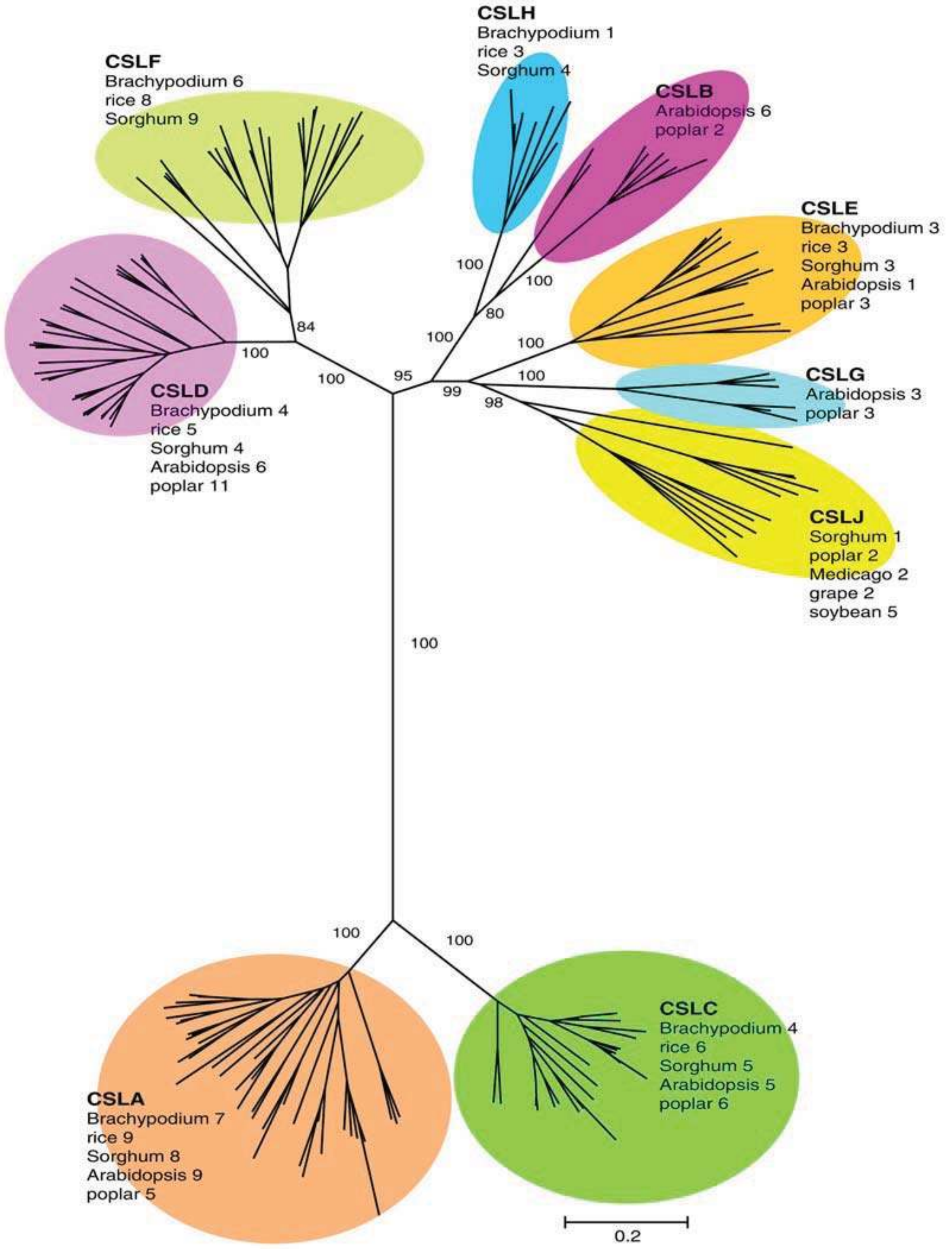
Gene family	General function	Gene models examined[1]	Gene models modified
Glycosyl hydrolase (GH)	cell wall modification	339	11
Pectin methylesterase Inhibitor (PMEI)	cell wall modification	38	0
Pectin methylesterase (PME)	cell wall modification	31	0
Laccase	cell wall modification	29	4
Glycosyl transferase (GT)	cell wall biosynthesis / polysaccharide biosynthesis	313	42
Putative Pectin MethylTransferase	cell wall biosynthesis (pectin)	23	0
Cellulose synthase-like (CSL)	cell wall biosynthesis (glucan)	25	7
DUF266 (putative glycosyl transferase)	cell wall biosynthesis (glucan)	19	0
Cellulose synthase	cell wall biosynthesis (glucan)	10	1
4-Coumarate:CoA ligase (4CL)	cell wall biosynthesis (lignin)	12	0
Phenylalanine ammonia lyase (PAL)	cell wall biosynthesis (lignin)	9	0
Cinnamoyl-CoA reductase (CCR)	cell wall biosynthesis (lignin)	9	0
Caffeoyl-CoA 3-O-methyltransferase (CCoAOMT)	cell wall biosynthesis (lignin)	8	0
Cinnamyl alcohol dehydrogenase (CAD)	cell wall biosynthesis (lignin)	7	0
Caffeic acid O-methyltransferase (COMT)	cell wall biosynthesis (lignin)	4	0
Ferulate 5-hydroxylase (F5H)	cell wall biosynthesis (lignin)	4	0
Hydroxycinnamoyl-CoA:shikimate/quinate hydroxycinnamoyltransferase (HCT (CST/CQT))	cell wall biosynthesis (lignin)	2	0
Trans-cinnamate 4-hydroxylase (C4H)	cell wall biosynthesis (lignin)	2	0
p-coumarate 3-hydroxylase (C3H)	cell wall biosynthesis (lignin)	1	0
RNA binding protein	RNA binding	282	141
NBS LRR	defense	178	0
bHLH transcription factor	transcription factor	149	3
AP2/ERF transcription factor	transcription factor	146	6
MYB transcription factor	transcription factor	109	28
NAC transcription factor	transcription factor	99	25
bZIP transcription factor	transcription factor	81	1
MYB-related transcription factor	transcription factor	71	2
WRKY transcription factor	transcription factor	71	8
MADS transcription factor	transcription factor	55	3

GRAS transcription factor	transcription factor	45	2
ABI3VP1 transcription factor	transcription factor	43	1
THX transcription factor	transcription factor	24	1
BEL1-LIKE homeodomain transcription factor	transcription factor	14	3
Homeodomain-Leucine Zipper II family protein	transcription factor	12	1
YABBY transcription factor	transcription factor	8	0
GARP transcription factor (G2-like transcription factor)	transcription factor	5	0
Homeobox transcription factors	transcription factor	16	3
Sulphate transporter	ion transporter	11	1
Autoinhibited Calcium P-type ATPase	ion transporter	10	1
Heavy Metal P-Type ATPase	ion transporter	9	2
Autoinhibited H+ P-type ATPase	ion transporter	9	4
Aminophospholipid P-type ATPase	ion transporter	9	3
ER- type Calcium/Manganese P-type ATPase	ion transporter	3	0
P5 P-type Atpase	ion transporter	1	0
Mitochondrial Molybdenum transporter	ion transporter	1	0
CrRLK1L	kinase	14	0
Phytochrome	photoreceptor	4	0
Homologous recombination protein	Recombination and DNA repair	16	0
Damage sensing and pre-processing recombination protein	Recombination and DNA repair	9	0
Accessory recombination protein	Recombination and DNA repair	7	0
Plastid specific recombination protein	Recombination and DNA repair	4	2
Non-Homologous recombination proteins	Recombination and DNA repair	3	0
Argonaute (AGO) Family	small RNA processing	15	0
Dicer-like (DCL) Family	small RNA processing	7	0
RNA-dependent RNA Polymerase (RDR) Family	small RNA processing	5	0
Prolamin	seed storage protein	15	3
Globulin	seed storage protein	14	1
Ha-like	seed storage protein	3	0
Starch Synthase	starch metabolism	10	0
Starch Branching Enzyme	starch metabolism	4	0
ADP-Glucose pyrophosphorylase, large subunit	starch metabolism	3	0
Isoamylase	starch metabolism	3	0
ADP-Glucose pyrophosphorylase, small subunit	starch metabolism	2	0
Pullulanase	starch metabolism	1	0
YUCCA-like flavin monooxygenase	auxin biosynthesis	23	0

PGP-like phosphoglycoprotein auxin transporter	auxin Transport	32	2
PINFORMED-Like Auxin Efflux Carrier	auxin Transport	10	4
Aux/LAX- Like Auxin Importer	auxin Transport	7	0
Cyclin	cell cycle	24	10
Cyclin-dependent kinase (CDK)	cell cycle	13	3
CKL	cell cycle	12	6
Anaphase promoting complex (APC)	cell cycle	11	2
Kip-related protein (KRP)	cell cycle	5	4
E2F	cell cycle	4	0
DP	cell cycle	3	1
DP-E2F–like (DEL)	cell cycle	2	0
Retinoblastoma (RB)	cell cycle	2	0
CDK subunit (CKS)	cell cycle	1	0
WEE1	cell cycle	1	1
VIN3 like (VIL)	chromatin modification	5	2
Extra sex combs like (ESCL)	chromatin modification	4	3
p55 like (p55L)	chromatin modification	4	1
Enhancer of zeste like (EZL)	chromatin modification	2	1
Suppressor of zeste 12 like (SUZL)	chromatin modification	2	2
Constans-like	circadian clock/flowering time	17	5
phosphatidylethanolamine-binding protein	circadian clock/flowering time	16	1
C2H2 transcription factor	circadian clock/flowering time	14	7
Apetala2 domain	circadian clock/flowering time	4	3
LOV-domain containing	circadian clock/flowering time	3	0
CCT-domain containing	circadian clock/flowering time	2	0
Gigantea	circadian clock/flowering time	1	1
heterochromatin protein1 family	circadian clock/flowering time	1	0
FLORICAULA/LEAFY-like	circadian clock/flowering time	1	0
Zea mays thick tassel dwarf1 (*TD1*) ortholog[2]	leucine-rich repeat receptor-like kinase	1	0
Zea mays ramosa2 (*RA2*) ortholog[2]	transcription factor	1	0
Zea mays teosintebranched1 (*TB1*) ortholog[2]	transcription factor	1	0
Zea mays YabbyA ortholog[2]	transcription factor	1	0
drought responsive genes from 11 families[2]	drought responsive gene	40	0
	total	2,755	369

[1] Includes eight genes manually added to the V1.0 annotation

[2] Genes from larger families selected for annotation based on putative function.



Supplementary Figure 9. Consensus neighbor-joining tree of the cellulose synthase-like (CSL) gene family based on 1,000 bootstrap trees. The number of genes found in the species examined is presented. For clarity, individual gene names are not shown. Note that the grasses have a similar distribution of family members with the exception of CSLJ, a family recently found in some grasses (wheat, barley, sorghum, maize) but not in *Brachypodium*, rice or Arabidopsis [28]. After identifying two poplar CSLJ genes we searched for additional dicot CSLJ genes in Medicago, soybean and grape and identified 9 genes that were added to the tree. Note that the sorghum and poplar gene models were not edited, so there may be additional CSL genes not represented because they were truncated or mis-annotated. Bootstrap support (% of 1,000) for the major branches is indicated.

Supplementary Table 10. Genes manually assigned to families. The Table shows genes that were specifically assigned to gene families and subfamilies, although these were not manually annotated.

Gene family	Number of genes	general function
Kinase (140 subfamilies)[1]	1,440	phosphorylation
RING	545	protein degradation
F-Box	489[2]	protein degradation
Bric-a-Brac/Tramtrack/ Broad Complex (BTB)	166[3]	protein degradation
U-box	70	protein degradation
26S	54	protein degradation
SKP1	16	protein degradation
Cullin	12	protein degradation
HECT	10	protein degradation
zf-Dof	27	transcription factor
auxin response factor (ARF)	24	hormone signaling
AUX/IAA	25	hormone signaling

[1] Since kinase family structure is not well defined in plants kinases were only assigned to subfamilies based on putative function.

[2] Includes 62 genes not included in the v1.0 annotation.

[3] Includes 67 genes not included in the v1.0 annotation.

Supplementary Table 11. Additional gene models identified in selected families. The v1.0 annotation contained fewer F-Box and BTB genes than expected based on previously sequenced genomes. To determine if additional genes were contained in the genome, but missed in the v1.0 annotation, we used domain scans to identify additional genes in these families. We also looked for additional genes in four smaller gene families to determine if missed genes were a systemic problem in the v1.0 annotation. We did not detect evidence for missing genes in these families.

Gene family	Gene models in V1.0 annotation	Additional gene models*	Total *Brachypodium* genes	Oryza	Sorghum	Arabidopsis	Populus
F-box	427	62	489	703	569	659	336
zf-Dof	27	0	27	30	29	36	42
Sucrose_synth	6	0	6	7	5	6	10
Auxin_resp	24	0	24	25	27	22	37
AUX_IAA	31	0	31	37	31	35	37
Bric-a-Brac/Tramtrack/ Broad Complex (BTB)	99	67	166	149	nd	80	nd

*All new models were supported by expression evidence.

Prediction of the *Brachypodium* secreted proteome

A comparative survey was conducted of the predicted secretome (proteins targeted to the secretory pathway) of *Brachypodium*, Arabidopsis and rice, to determine whether the substantial differences between grass and dicot cell wall architectures [38] might be mirrored in distinctive populations of proteins that enter the secretory pathway. Three prediction methods were used to detect the presence of N-terminal signal peptides (SP) in the predicted proteomes of each species: TargetP (www.cbs.dtu.dk/services/TargetP) and SignalP (www.cbs.dtu.dk/services/SignalP) neural network (NN) or hidden Markov model (HMM). SignalP NN, which gave the lowest inter-species variation on a per-genome percentage (Supplementary Table 12), was selected as generating the most accurate prediction because it had the smallest proportions of apparent false positive or negative predictions following manual inspection (not shown).

Supplementary Table 12. Computational prediction of genes from Arabidopsis, *Brachypodium* and rice encoding proteins targeted to the secretory pathway. The total number of proteins/unigenes used in the search for each species is given in parentheses underneath each species.

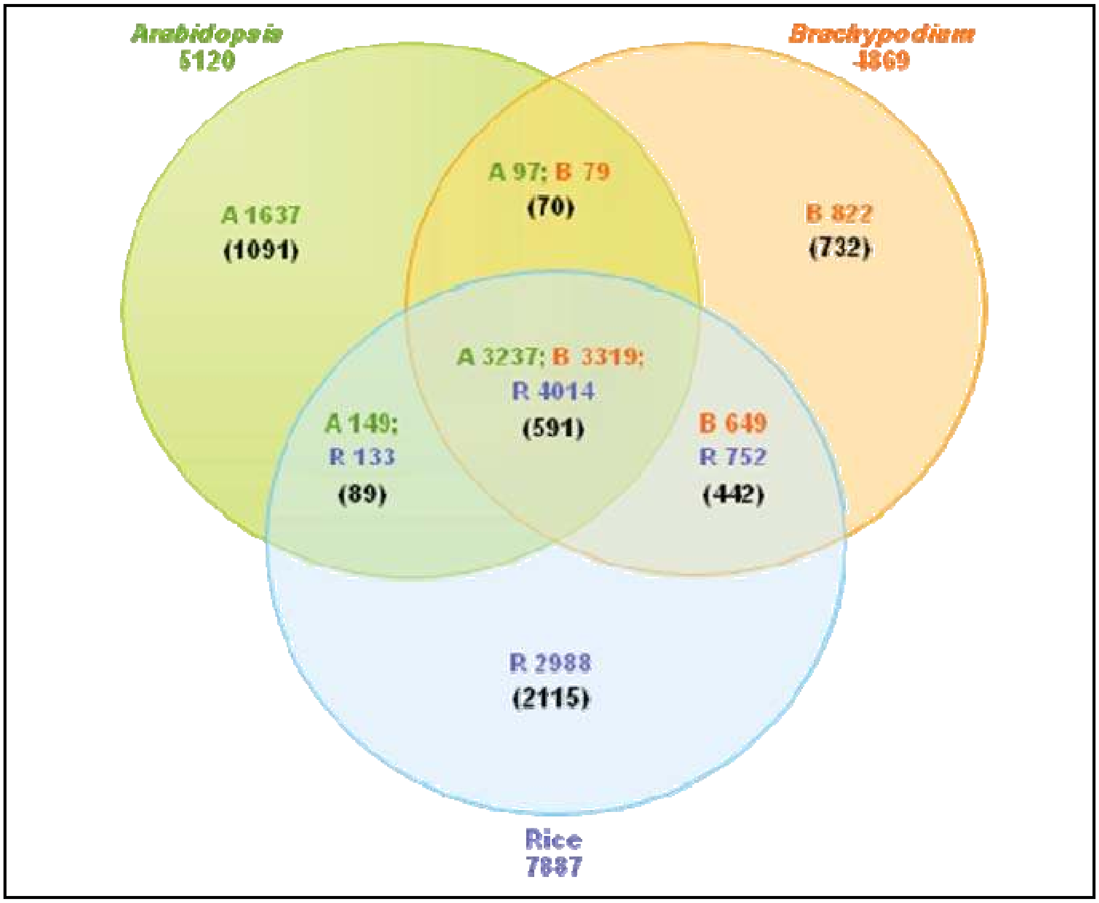
Program	Arabidopsis (27,011)	*Brachypodium* (25,532)	Rice (55,807)
TargetP	5,338 (19.8%)	4,272 (16.8%)	6,921 (12.4%)
SignalP HMM	6,064 (22.5%)	7,542 (29.7%)	12,966 (23.2%)
SignalP NN	5,120 (19.0%)	4,869 (19.1%)	7,887 (14.1%)

The secreted proteins predicted by SignalP NN from *Brachypodium*, Arabidopsis (TAIR8 version), and rice (TIGR v6) were clustered using the homolog clustering algorithm TribeMCL [39]. A total of 3,319 (68.2%) *Brachypodium* genes encoding SP-containing proteins were shared among all three species, 3,398 (69.8%) with Arabidopsis, 3,968 (81.5%) with rice and 4,047 (83.1%) with at least one of the other two species (Supplementary Figure 10).

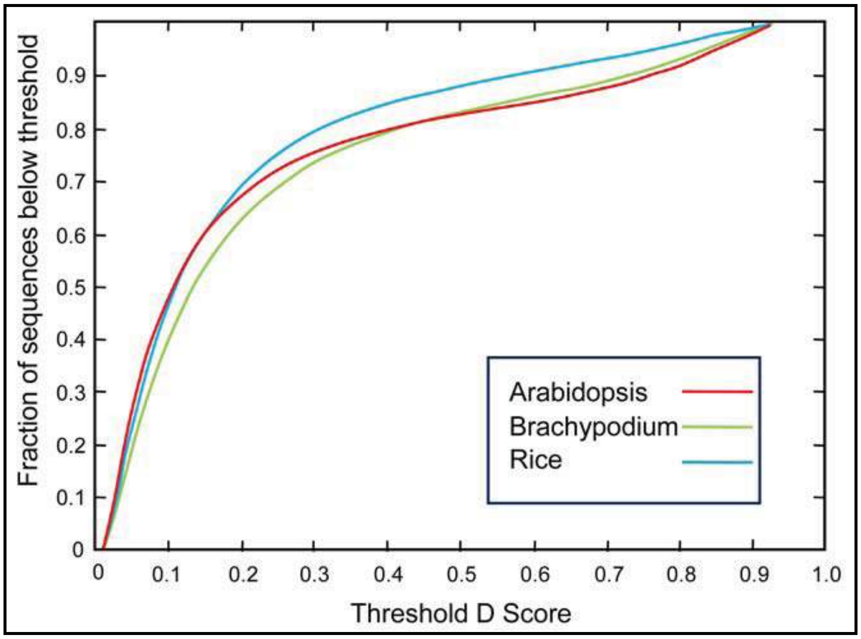
This analysis identified some substantial differences in the relative sizes of some specific secreted families in dicots and grasses, particularly in the distribution of cell wall metabolism genes (see Supplementary Table 13). 26 pectate lyase genes were identified in Arabidopsis, 29 in poplar, but only 7 in *Brachypodium*, 12 in rice and 10 in sorghum, consistent with the low pectin levels found in grass cell walls compared to dicots [38]. Conversely, members of the superfamily of expansins, which play a major role in cell-wall loosening [40], are more abundant in monocots (61 in *Brachypodium*, 58 in rice, and 88 in sorghum) than in dicots (35 in Arabidopsis and 43 in poplar). In grass species, the size of the beta-expansin subgroup is particularly large. Some beta-expansins are also known as group 1 grass pollen allergens that are thought to promote wall loosening and facilitate pollen tube growth in the stylar tract, while others are also expressed in vegetative tissues. This suggests either that expansins have more than one substrate or activity in Type II grass walls, or they may have additional biological functions.

Glycosyl hydrolase family 5 (GH family 5) proteins are known to have mannan hydrolase and transglycosylase activity (www.cazy.org) [41] in plants, and likely contribute to wall remodeling in various developmental processes, including cell expansion and fruit ripening [37]. We identified 10 GH5 genes in *Brachypodium* and 17 each in rice and sorghum belonging to three subfamilies of secreted proteins (Sec family 515, 1219 and 2860), compared with 13 in Arabidopsis and 25 in poplar that lacked members of the Sec family 2860. This suggests that the secreted proteins in Sec family 2860 may contribute to the monocot-specific cell wall metabolism. Interestingly mannans are typically minor components of monocot walls, but it has been suggested that monocot GH5 isozymes may act as hetero-transglycosylases, which could explain their relatively high abundance. This is also the case with some xyloglucan endotransglucosylase-hydrolases (XTHs) which can catalyze heterotransglycosylation between xyloglucan and other polysaccharides [37]. The activities of plant GH5s are still poorly understood.

We also determined that a subfamily of dirigent proteins, which are proposed to be involved in the formation of lignans and the control of phenoxy radical-radical coupling reactions, are more abundant in monocots (49 in *Brachypodium*, 72 in rice, and 55 in sorghum) than in dicots (23 in Arabidopsis 38 in poplar). They are likely to function in the synthesis of specific lignans, but this has yet to be explored.



Supplementary Figure 10. Venn diagram of genes carrying a predicted signal peptide between Arabidopsis (A), *Brachypodium* (B) and rice (R). The number of *Brachypodium* signal peptide-containing protein genes is similar to that of Arabidopsis. Numbers in parentheses indicate the number of ABR protein families.



Supplementary Figure 11. The secreted proteomes of Arabidopsis, rice and *Brachypodium*. N- terminal signal peptides (SP) were predicted using signal P NN (www.cbs.dtu.dk/services/SignalP). The distribution of D probability scores was very similar for *Brachypodium* and Arabidopsis, indicating the start codons of genes were accurately predicted in *Brachypodium*.

Supplementary Table 13. Examples of signal peptide-containing protein families from *Brachypodium* and rice with differential abundance in *Brachypodium* and Arabidopsis. Note that with the exception of the dirigent protein subfamily, *Brachypodium* and rice show similar differences with respect to Arabidopsis.

Sec fam	*Species*	*Number of genes*	*Number of SP-containing genes*	*Annotation*
63	*Arabidopsis*	*35*	*28*	*Expansin*
	Brachypodium	*61*	*58*	
	Rice	*58*	*56*	
208	*Arabidopsis*	*9*	*6*	*Glycosyltransferase family 37* *(putative fucosyltransferases)*
	Brachypodium	*16*	*3*	
	Rice	*21*	*1*	
216	*Arabidopsis*	*26*	*23*	*Pectate lyase*
	Brachypodium	*7*	*2*	
	Rice	*12*	*8*	
524	*Arabidopsis*	*17*	*17*	*Subfamily of invertase/pectin methylesterase inhibitor proteins*
	Brachypodium	*2*	*2*	
	Rice	*4*	*4*	
582	*Arabidopsis*	*1*	*1*	*Glycosyltransferase family 31, Group F* *(putative galactosyltransferase)*
	Brachypodium	*10*	*7*	
	Rice	*10*	*7*	
1029	*Arabidopsis*	*5*	*5*	*Subfamily of dirigent proteins*
	Brachypodium	*0*	*0*	
	Rice	*8*	*7*	

Repeats Analysis

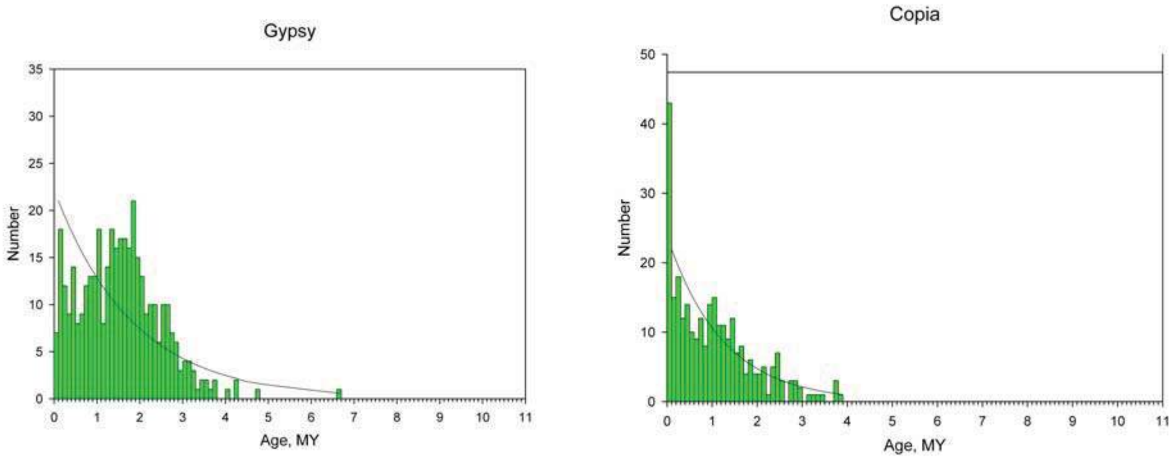
LTR retrotransposons

De novo searches for LTR retrotransposons were performed with LTR_STRUCT and LTR_HARVEST [42]. Duplicates were removed with CD_HIT and the resulting LTR pairs were checked with DOTTYP from the EMBOSS package and by visual inspection. This identified 891 full-length LTR retrotransposon candidate sequences that were assessed for typical retrotransposon protein domains (GAG, AP, IN, RT) by an HMMer (http://hmmer.janelia.org) search against respective PFAM HMM models and against the REPEATMASKER libraries. Searches were also made against PTREP and PFAM using EXONERATE v.2.2. Complex nests were removed from the library. 690 (78%) of the candidate sequences remained after a quality check and overlap removal. The main quality criteria were the existence of at least one typical retrotransposon protein domain and a simple sequence and tandem repeat content<=35%. Superfamily membership was assigned by protein signature. The *Gypsy* superfamily (AP-RT-IN) predominates throughout the *Brachypodium* genome, where it is the most abundant group of transposable elements, contributing 55.4% of the total retrotransposons in a total of 19 clusters defined by the first 24 nt of the LTR, compared with 40.8% for the *Copia* superfamily in a total of 44 clusters. The *Gypsy* superfamily contributes 70.6% of the intact LTR retrotransposons and covers16.1% of the genome, or 3.3 times more than *Copia*. Only 3.8% of the intact elements, forming 9 clusters, could not be placed in a superfamily. *Brachypodium* displays appreciable chromosome-to-chromosome differences in the distribution of LTR retrotransposons. Chromosome 5 is richest, with 28.3% coverage by retrotransposons (intact elements, solo LTRs, fragments), and chromosome 1 the poorest, with only 20.3%. Chromosome 4 is deficient in *Gypsy* elements (2.34 times less abundant), whereas chromosome 5 is enriched (2.9 times more abundant). Chromosome 5 also has the youngest Gypsy elements (1.37 MY vs. 1.54 – 1.64 MY for the others). Chromosome 4 has 18 of the 52 intact elements younger than 0.1 MY, whereas chromosome 5 has only four.

The set of 690 high-quality LTR retrotransposons were added to mipsREdat (mips.gsf.de/proj/plant/webapp/recat/), a plant repeat element database, and used for homology based repeat masking and annotation. Clustering of LTR retrotransposons was based on the first 25 nt of the 5' UTR following alignment with CLUSTALW and hand editing with the aid of the GENEIOUS package (htpp://www.geneous.com). Global pairwise alignments were for the LTRs of each element constructed with NEEDLE from the EMBOSS package. The insertion age of full length LTR-retrotransposons was determined from the evolutionary distances between 5' and 3' solo LTRs, which were calculated with FDNADIST of EMBOSS. For the conversion of distance to insertion age, a substitution rate of 1.3E-8 mutations per site per year was used [43]. Half-life (t1/2) was estimated by fitting an exponential decay curve, using the formula y=a*2exp-(t/t1/2) by least-squares individually to the numbers of *Copia* and *Gypsy* intact elements, summed for each bin of 0.1 MY, as previously described [44].

A total of 1,814 solo LTRs was identified in *Brachypodium* by similarity search to the full-length elements and by structural analysis. These represent only 0.25% of the genome. Assuming that each solo LTR (average length 379 bp) was derived from an intact element of 10 kb, a minimum of 17.4 Mb is predicted to have been lost from the genome by LTR : LTR recombination. This represents 2.7 times the current genomic coverage by intact elements (6.47 Mb), but ignores possible recombinations between solo LTRs subsequent to their production and hence may be an underestimate. The *Gypsy* solo LTRs (1,122) are 1.6-fold more abundant than the *Copia* solo LTRs (689), similar to the relative abundance of intact *Gypsy* elements (1.36). Of all the intact elements in the *Brachypodium* genome*,* 483 (69.8%) have no related solo LTRs, and 81 have one. The Bd3_RLG_17 element (0.69 MY old) has 645 related solo LTRs and Bd3_RLC_6 (0.45 MY old) has 263. Both elements are widespread in the Triticeae. The ratio of the number of solo LTRs to the age of the related intact elements indicates the propensity to form solo LTRs. The three elements in the genome with the highest value for this measure include the Bd2_RLC_14 element, which belongs to the *Angela* – *BARE* – *Wis* family and is 20,769 years old, yet has 35 solo LTRs associated with it. The Bd4_RLC_10 element is similar to SC-7 of rice, is less than 20,000 years old, and has two solo LTRs. The recent activity of the *Angela* – *BARE* – *Wis* family members in the *Brachypodium* genome is further evidence for the role of retrotransposon loss through recombination as a way of controlling genome size expansion.

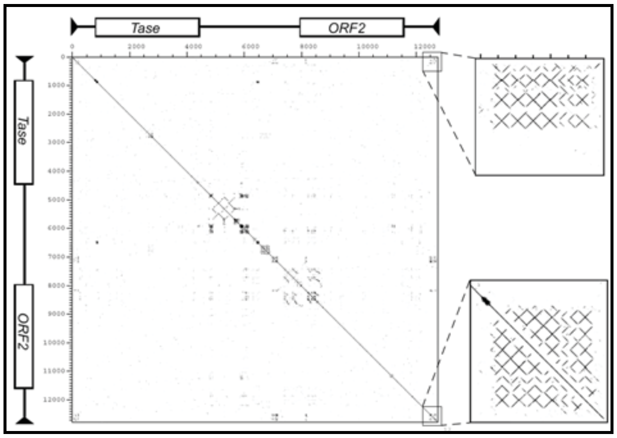
The distribution of solo LTRs between the chromosomes is strikingly different. While the chromosomes have on average 362 solo LTRs each, chromosome 5 has only 73, whereas chromosome 3 has 1,016. Chromosome 5 contains one solo LTR per 389 kb, whereas chromosome 3, also the richest by this measure, has one per 239 kb. Chromosome 3 is also home to the two most abundant sets of solo LTRs in the genome, Bd3_RLC_17 and Bd3_RLC_6. Solo LTRs cannot be mobilized, and remain at the loci where they are produced by recombination. Hence, the ratio of solo LTRs to intact LTR retrotransposons gives an indication of the relative rates of repetitive DNA gain through integration of new elements and loss through recombination. Whereas the genome as a whole has a ratio of 2.6 solo LTRs to each intact elements, chromosome 5 has a ratio of only 0.89, and chromosome 3 has 6.96; the others have ratios between 1.23 and 1.73. When taken together with the number and age of the full-length LTR retrotransposons, these data suggest that chromosome 5 is gaining retrotransposons by replication and losing comparatively few by recombination.



Supplementary Figure 12. Retroelement family ages in the *Brachypodium* genome. The age distribution and frequency of intact *Copia* and *Gypsy* LTR retrotransposons (green bars) grouped in age classes of 0.1 MY. Fitted exponential decay curves for the half-life of intact elements are shown. Half-life for *Gypsy* elements, 1.265 MY; for *Copia* elements, 0.859 MY.

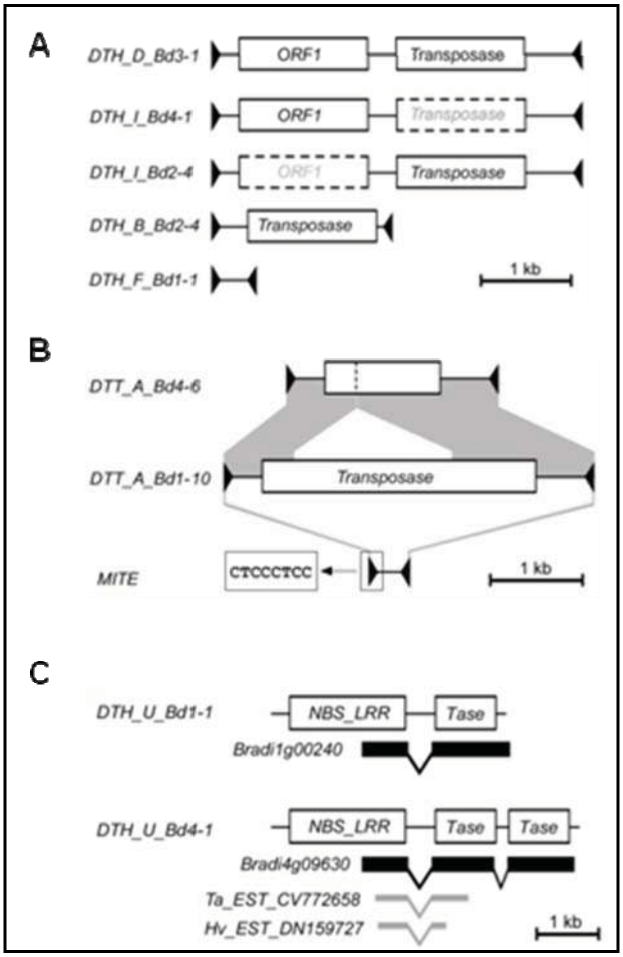
Identification and characterization of Class 2 transposons

Candidates for *CACTA* transposons were identified with a Perl program that searched the genome in sliding windows for CACTA…TAGTG motifs that are separated by 8-12 kb and flanked by a 3 bp target site duplication (TSD). This produced many false positives as such patterns can occur by chance. In a second step, candidate CACTAs were screened for the presence of a transposase and the ORF2 by BLASTX against the protein division of TREP (http://wheat.pw.usda.gov/ITMI/Repeats/). Those that produced blast hits were manually checked by DotPlot for the presence of full-length ORF and intact ends, which typically contain arrays of direct and inverted repeats (Supplementary Figure 13). Once a full-length element was identified, all similar copies were extracted from the candidate set and a consensus was constructed.



Supplementary Figure 13. Characteristics of a CACTA transposon visualized by DotPlot. The two insets show the typical sub-terminal arrays of direct and inverted repeats.

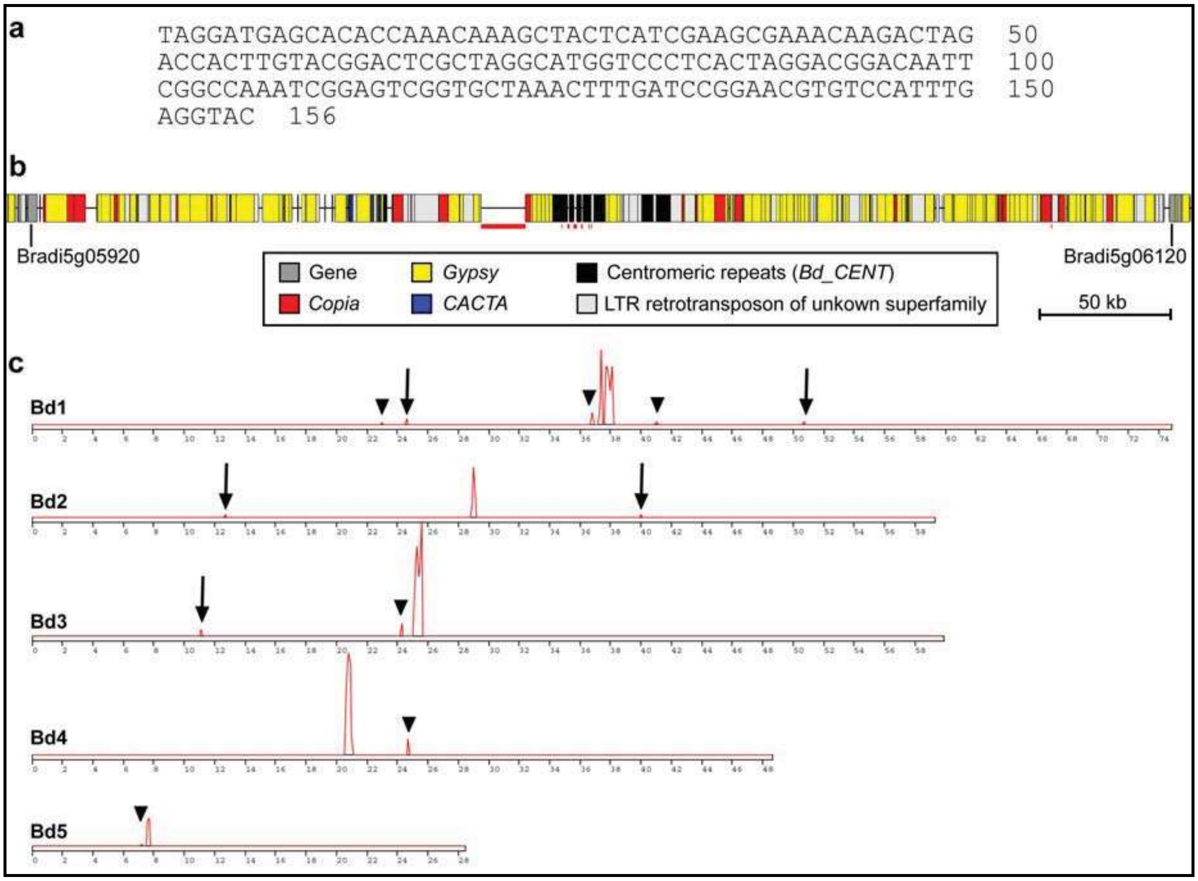
Candidates for autonomous elements of the *Mariner* and *Harbinger* superfamilies were identified by TBLASTN of known elements against the whole genome. All regions that produced significant hits (E-values <10^{-10}) were excised with 5-10 kb of flanking regions with the help of a Perl program. Terminal inverted repeats were identified by DotPlot. Consensus sequences for families with sufficiently high copy numbers were produced as described above. TIRs of *hAT* elements were identified by NCBI-BLAST2 of known TIRs. Complete elements were verified manually by identification of the TSD. To identify non-autonomous *hAT* elements, full-length elements were used for RepeatMasker analysis of the whole genome. Candidate full-length elements were extracted with their flanking regions with a Perl script. Complete elements were verified manually by identification of TSDs. RepeatMasker was reiterated until no new full-length TEs were identified. The combined dataset was then used for RepeatMasker analysis of the whole genome to identify partial elements. *Mutator* elements were identified in two ways. First as above by identification of CDS and by screening the genome for large inverted repeats that are flanked by a 9 bp TSD. Candidates for Miniature inverted-repeat transposable elements (MITES) were also detected based on their inverted repeat structure. For *Stowaway* MITEs, the typical CTCCCTCC termini were used as an additional criterion. All Perl scripts that were written for the identification of Class 2 elements are available upon request.



Supplementary Figure 14. DNA transposon structures in *Brachypodium*. *A*). *The typical Harbinger (DTH) autonomous element (top) has two ORFs. Semi-autonomous* elements have one intact and one degenerate ORF (dashed lines). Some families (e.g. *DTH_B*) contain only one or no ORF at all (e.g. *DTH_F*) and probably recruit the gene products of other *Harbinger* families for transposition. **B)**. Recent and ancient deletion derivatives. The recent deletion derivative (top) shows strong sequence homology with its Mother element (middle) and the deletion breakpoint (dashed line) can be determined precisely. In the ancient deletion derivative (MITE, bottom) only the very terminal few bp are conserved. **C)**. Fusion of an NBS-LRR gene to a *Harbinger U* transposase gene. The chimeric gene model is indicated as a black bar with introns as bent lines connecting exons. The novel gene is conserved in Triticeae, shown by the ESTs from wheat and barley (grey bars). Tase, fused transposase gene.

Brachypodium centromeres

The consensus sequence of the *Brachypodium* centromeric repeat (BdCENT) is 156 bp long (Supplementary Figure 15A), very similar in size (but not in sequence) to those of rice, sorghum and Arabidopsis (155 bp, 137 bp and 159 bp, respectively). The centromere of *Brachypodium* chromosome 5 is essentially complete, with one central join (Supplementary Figure 15B). It is ~45 kb and consists of two BdCENT arrays, one with >88 (containing 6 sequence gaps) and one with 20 repeat units. The other centromeres are approximately 162 kb to 798 kb and contain up to 1300 repeat units. These are minimal numbers as all centromeres contain sequence gaps. BdCENT arrays are occasionally interspersed with large blocks of LTR retrotransposons. Eleven additional regions contained 1 to 49 BdCENT units; five correspond to chromosome fusion points (Figure 4A), demonstrating that chromosomes inserted precisely into the centromeres of others during grass chromosome evolution (Supplementary Figure 15C and Figure 4A). All centromeres are flanked by gene-poor regions with high numbers of LTR retrotransposons, almost exclusively of the *Gypsy* superfamily. Within 300 kb of all five centromeres, only 54 genes were found, all of which were non-collinear in rice and sorghum. Bd1 contains a set of 10 genes and Bd2 contains one gene inside the centromeric repeat cluster. The other centromeres are free of genes.



```
TAGGATGAGCACACCAAACAAAGCTACTCATCGAAGCGAAACAAGACTAG   50
ACCACTTGTACGGACTCGCTAGGCATGGTCCCTCACTAGGACGGACAATT   100
CGGCCAAATCGGAGTCGGTGCTAAACTTTGATCCGGAACGTGTCCATTTG   150
AGGTAC  156
```

a
b
Bradi5g05920
Bradi5g06120
Gene
Copia
Gypsy
CACTA
Centromeric repeats (Bd_CENT)
LTR retrotransposon of unkown superfamily
50 kb
c
Bd1
Bd2
Bd3
Bd4
Bd5

Supplementary Figure 15. Sequence organisation of *Brachypodium* centromeres. a. Consensus sequence of the *Brachypodium* centromeric repeat unit (*Bd_CENT*). b. Map of the centromeric region of *Brachypodium* chromosome 5 (Bd5). Centromeric and pericentromeric regions up to the first flanking genes are shown. Sequence gaps are indicated as red bars underneath the map. c. Distribution of *Bd_CENT* repeats along *Brachypodium* chromosomes. Occurrences of *Bd_CENT* repeats outside of the centromers are indicated with arrows and arrowheads. Arrows indicate *Bd_CENT* arrays that correspond to chromosome fusion points.

Repeat data integration

The integration of transposon data from different expert groups 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 are caused by highly similar regions shared by different transposons or by composite elements e.g. LTR retrotransposons with MITE inserts. Such annotation overlaps were handled using a priority based approach. High confidence expert annotations were assigned first, with a higher priority on young full length elements, which still possess target site duplications. Overlapping elements with lower priority were either truncated, fragmented or skipped, depending on adjustable parameters for overlap percent and minimum length. The assignment order within one priority group was defined by descending homology score or element length. For *Brachypodium* all elements overlapping > 80% of their length to higher priority elements were removed. Elements overlapping by ≤80% were truncated or split, if the remaining length exceeded 49 bp. In the first step overlaps within each of the 10 different annotations were removed. The following priority order was used in the next step: 1. Mariner (DTT) 2. Pif-Harbinger (DTH) 3. tourist_MITEs (DTH) 4. stowaway_MITEs 5. CACTA (DTC) (DTT) 6. hAT (DTA) 7. Full length LTR-retrotransposons (RLX, RLG, RLC) 8. Helitrons (DHH), 9. Mutator (DTM) 10. RIX (LINEs), 11. LTR-retrotransposon fragments. Step 1-7 were applied in 2 iterations, first with full length elements still having target site duplications and second with the remaining elements of the respective group. The resulting transposon annotation was named Brachy_transposons_v2.2. A summary of the annotated transposon content of *Brachypodium* is shown in Supplementary Table 14, and features of DNA transposons are shown in Supplementary Figure 14.

Supplementary Table 14. *Brachypodium* transposable element content. The table summarizes the annotation of full length elements and transposon fragments that were classified according to [45].

	families	copies	% copy number	Mb	avg length bp	% of TE bp	% of genome
Mobile Element (-)		80,049	100.00	76.091	951	100.00	28.10
Class I: Retroelement (RXX)		**50,419**	**62.99**	**63.168**	**1,253**	**83.02**	**23.33**
LTR Retrotransposon		47,274	59.06	57.908	1,225	76.10	21.39
full length		690	0.861972	6.468	9,373	8.4999	2.3885036
solo		1,814	2.266112	0.685	378	0.900762	0.2531174
Ty1/copia (RLC)	44	12,426	15.52	13.149	1,058	17.28	4.86
full length		282	0.35	1.900	6,737	2.50	0.70
solo		689	0.86	0.332	482	0.44	0.12
Ty3/gypsy (RLG)	19	32,978	41.20	43.464	1,318	57.12	16.05
full length		382	0.48	4.358	11,408	5.73	1.61
solo		1,122	1.40	0.352	313	0.46	0.13
unclassified LTR (RLX)	9	1,870	2.34	1.295	693	1.70	0.48
full length		26	0.03	0.210	8,074	0.28	0.08
solo		3	0.004	0.002	567	0.002	0.001
non-LTR Retrotransposon (RXX)		3,145	3.93	5.259	1,672	6.91	1.94
LINE (RIX)		3,145	3.93	5.259	1,672	6.91	1.94
Class II: DNA Transposon (DXX)		**29,630**	**37.01**	**12.924**	**436**	**16.98**	**4.77**
Superfamily (DTX)		5,947	7.43	9.564	1,608	12.57	3.53
CACTA (DTC)	14	1,523	1.90	5.899	3,873	7.75	2.18
HAT (DTA)	56	658	0.82	0.644	978	0.85	0.24
Mutator (DTM)	65	2,854	3.57	1.710	599	2.25	0.63
Tc1/Mariner (DTT)	8	50	0.06	0.177	3,542	0.23	0.07
PIF/Harbinger (DTH)	24	862	1.08	1.135	1,316	1.49	0.42
MITE (DXX)		23,563	29.44	2.869	122	3.77	1.06
Stowaway (DTT)	21	20,994	26.23	2.394	114	3.15	0.88
Tourist (DTH)	19	2,569	3.21	0.475	185	0.62	0.18
Helitron (DHH)	48	120	0.15	0.491	4,089	0.64	0.18

Simple Sequence Repeats

SSRs were located using SSRLocator [46]. It was configured to locate perfect, imperfect and composite SSRs [47] ,Class I (≥ 20 bp) and Class II (≥ 12 and < 20 bp) repeats [48], and classify repeats according to length: 12x monomer, 6x dimer, and 4x trimer repeats and 3x tetramer, pentamer, and hexamer repeats. In this analysis, monomer to hexamer repeats were considered, according to [49,50]. SSRs were integrated with gene annotations and classified as intronic, exonic or intergenic. The distribution of simple sequence repeats (mono- up to hexamers) are shown in Supplementary Table 15. In *Brachypodium* trimers (37.6%) and tetramers (32.7%) are the most abundant (70.3%), compared to Arabidopsis and rice where they are rarer (50.0% and 62.0% respectively). Short repeats (Class II) predominate over long repeat (Class I) loci respectively, totalling 91,434 (93.3%) and 6,593 (6.7%). Class II predominates for all types of repeats in terms of numbers of loci, numbers of repeats, and total length in base pairs. G/C monomer motifs predominate when all (62.5%) or when only Class I (90.1%) repeats are assessed. For dimers, AG/GA, AT/TA and CT/TC predominate when all (72.9%) or only Class I (82.8%), were assessed. G/C-rich trimers, independent of sequence arrangement motifs, predominate (35%). For tetramer, pentamer and hexamer motifs, there was no apparent predominance of a given motif. SSRs are overwhelmingly present in intergenic (88.0%) regions when compared to exonic (6.2%) and intronic (5.8%) regions. Class I SSRs show a similar trend, except for the preference for intronic (2-fold higher) compared to exonic regions. In general, trimers and hexamers predominate in exons (92.0%) while trimers and tetramers predominate in introns (66.1%) and intergenic regions (69.2%). Class I SSRs show similar results for exons, but dimers and monomers increase significantly when introns and intergenic regions are assessed.

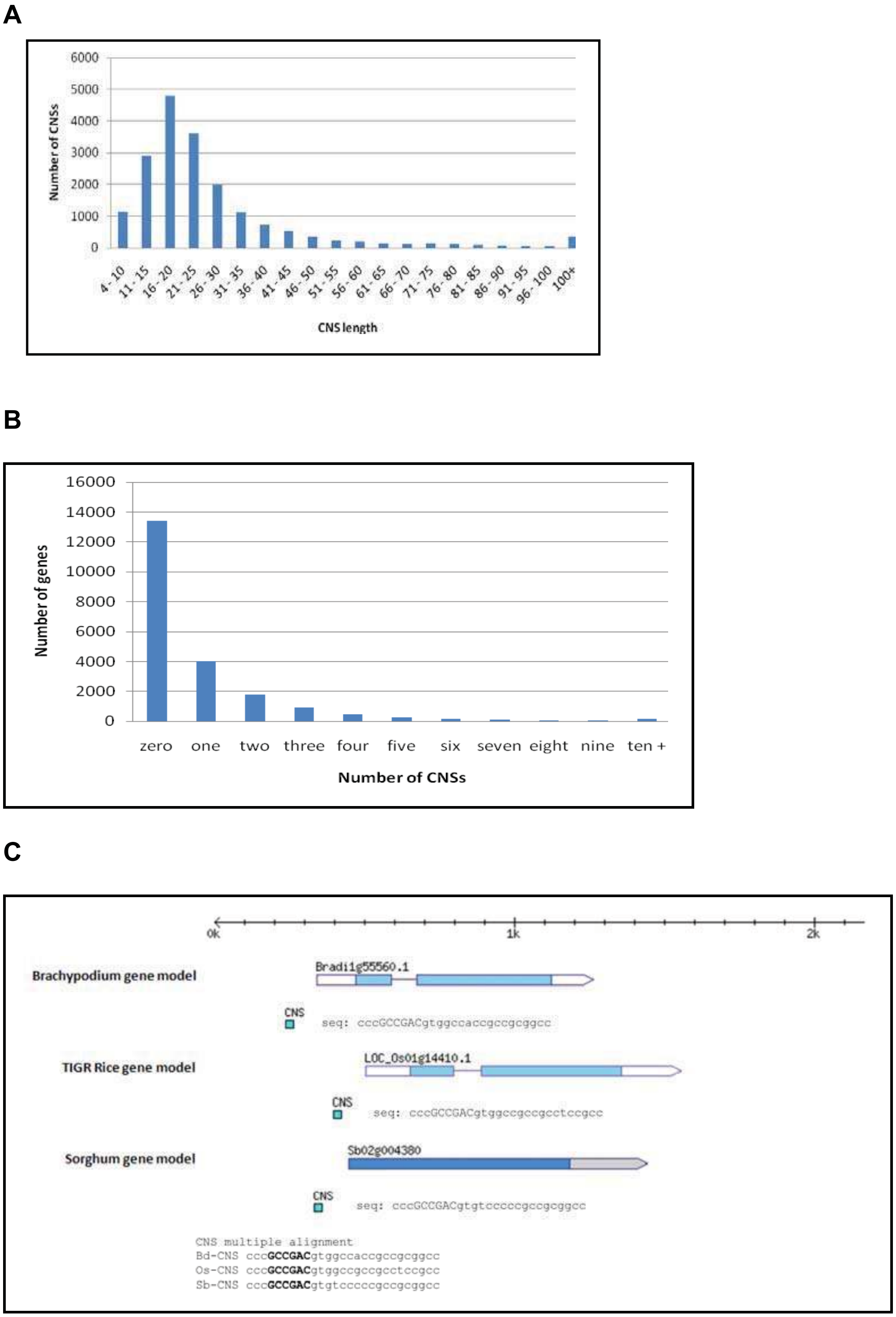
Supplementary Table 15. Summary of simple sequence repeat (SSR) types and numbers in the *Brachypodium* genome.

Type	**Class**	**Total**	**Total**	**Total**	**Average**	**Repeat**
		Loci	**Repeats**	**Length (bp)**	**Length (bp)**	**Numbers**
			(nº repeats)	**(nº repeats * type)**	**(Total length / Total loci)**	
Monomers	I	789	18,344	18,344	23.2	>= 20
	II	7,207	100,883	100,883	14.0	>= 12 and <= 19
	total	**7,996**	**119,227**	**119,227**	**14.9**	
Dimers	I	1,676	26,102	52,204	31.1	>= 10
	II	7,689	52,361	104,722	13.6	>= 6 and <= 9
	total	**9,365**	**78,463**	**156,926**	**16.8**	
Trimers	I	1,656	15,349	46,047	27.8	>= 7
	II	35,236	152,107	456,321	13.0	>= 4 and <= 6
	total	**36,892**	**167,456**	**502,368**	**13.6**	
Tetramers	I	979	5,990	23,960	24.5	>= 5
	II	31,068	96,378	385,512	12.4	>= 3 and <= 4
	total	**32,047**	**102,368**	**409,472**	**12.8**	
Pentamers	I	1,007	4,349	21,745	21.6	>= 4
	II	6,922	20,766	103,830	15.0	= 3
	total	**7,929**	**25,115**	**125,575**	**15.8**	
Hexamers	I	486	2,091	12,546	25.8	>= 4
	II	3,312	9,936	59,616	18.0	= 3
		3,798	**12,027**	**72,162**	**19.0**	
Total/Average		**98,027**	**504,656**	**1,385,730**	**14.1**	

					Average	**ssr/mb**
Occurrence	**Repeat Total**	**%**	**bp total**	**%**	**Number repeats**	
Class I	6,593	6.7	174,846	12.6	26.5	24
Class II	91,434	93.3	1,210,884	87.4	13.2	334
Total	**98,027.0**		**1,385,730**		**14.1**	

Conserved Non-coding Sequences

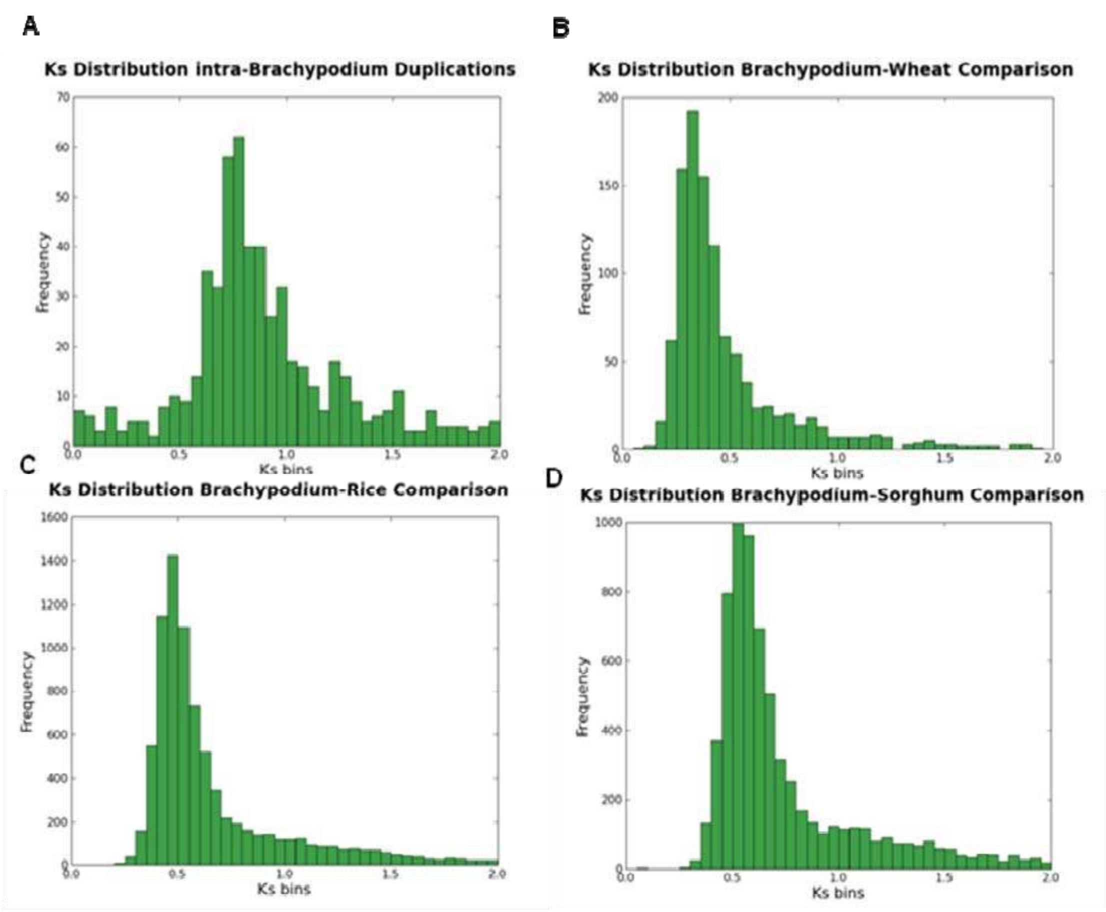
The predicted proteomes of *Brachypodium* (v1.0), sorghum (v1.4) and rice (TIGR v5) were used as input into OrthoMCL v1.4 [51] to determine putative rice and sorghum orthologs of each *Brachypodium* gene. 21,480 genes were included in orthologous sets. The genome sequence of orthologs spanning the mid-points of adjacent genes was extracted. Exons were masked and bl2seq v2.2.18 [52] was used to run pair-wise comparisons between the *Brachypodium* sequence and each of its rice and sorghum orthologs using settings designed to identify short conserved sequences as previously described [53]. A spike sequence was used to reduce the noise in the BLAST results [54]. The resulting HSPs were post-processed to identify regions on the *Brachypodium* sequence that were covered by both a *Brachypodium*-rice HSP and a *Brachypodium*-sorghum HSP. Only HSPs having a percentage identity of 85% or higher were included in this step and overlapping regions of less than 4bp were excluded. Using these stringent criteria we identified 18,664 sequence regions that are conserved between orthologous genes in *Brachypodium*, sorghum and rice, 11,328 of these are syntenic in the three genomes (true CNSs) and 7,336 are conserved but non-syntenic. These conserved sequences in the *Brachypodium* genome have lengths ranging from 4 to 2255 nucleotides (Supplementary Figure 16A: mean length 28 bp, median length 21 bp, 0.87 CNS per gene). The majority of *Brachypodium* genes have no CNS, 4008 genes have one CNS and 4042 have two or more CNSs including 153 genes that have more than 10 CNS each (Supplementary Figure S16B). We identified potentially functional motifs in some of these CNSs, such as DRE/CRT drought response motifs [55] (Supplementary Figure 16C).



Supplementary Figure 16. Conserved non-coding sequences in *Brachypodium* **A.** Distribution of CNS lengths**. B**. Distribution of the number of CNS per gene**. C.** CNS upstream of orthologous genes in *Brachypodium*, rice and sorghum. The multiple sequence alignment shows the core DRE/CRT (dehydration-responsive element/C-repeat) cis-acting element in bold. Expression of the rice gene is increased in response to drought [56].

Ks analysis of whole genome ortholog comparisons

Orthologs of *Brachypodium* genes were determined in rice (TIGR5) and sorghum (v1.4) genes as described in Supplementary Figure 6. For wheat orthologs, all possible three-frame translations from ESTs were determined and the best matching open reading frame was determined by a blastp comparison against the *Brachypodium* orthologous protein sequence. Nucleotide sequences were trimmed according to the blastp alignment to fit deduced open reading frames. Smith-Waterman alignments (EMBOSS package) [57] were generated for each orthologous protein pair and transformed to pairwise codon based alignments. Codeml of the PAML package [58] using the F3x4 model was applied to estimate Ka and Ks by maximum-likelihood and by the method of [59].



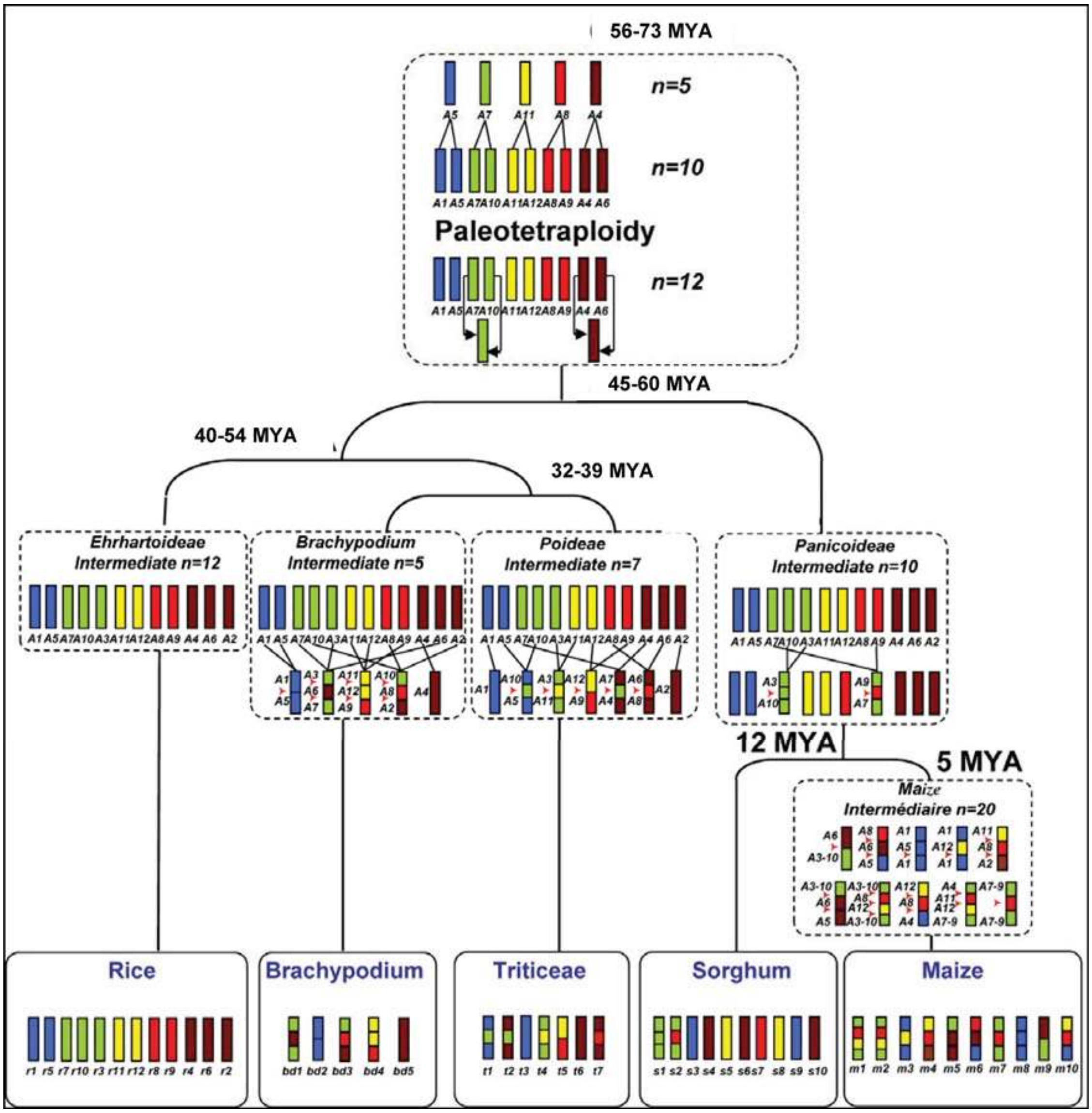
Supplementary Figure 17. Ks Distributions of intra-genomic *Brachypodium* duplications and *Brachypodium*, sorghum, rice and wheat orthologous genes. The charts show Ks values derived by the maximum-likelihood method [58]. The bin size of Ks values is 0.05. Note that the wheat distributions are based on translated EST data and may overestimate mean Ks due to higher sequencing errors in ESTs. A. Whole genome duplications in *Brachypodium*. B. *Brachypodium*- wheat ESTs. C. *Brachypodium*- rice. D. *Brachypodium*- sorghum.

Supplementary Table 16. Mean Ks and divergence times for *Brachypodium* versus several monocot species. Mean Ks and divergence times were obtained from the Ks distributions of syntenic pairs between *Brachypodium* and the monocot species listed in the first column. NG (Nej-Gojobori), ML (Maximum-Likelihood). Divergence times were calculated assuming a λ=6.1x10^{-9} (mean of 5.1-7.1x10^{-9}) [60]. Ks estimates for wheat may be overestimated as they are based on EST data. Figure 3A shows a cartoon of the divergence times of the different monocot groups estimated from this analysis.

Species	**Method**	**Mean Ks**	**Divergence time [10^7 a]**
***Brachypodium distachyon*, internal duplications**	NG	0.6842	5.61
	ML	0.8894	7.29
Triticum aestivum* (Wheat*)	NG	0.3956	3.24
	ML	0.4779	3.92
Oryza sativa ssp japonica **(Rice*)***	NG	0.4950	4.06
	ML	0.6581	5.39
Sorghum bicolor **(Sorghum)**	NG	0.5500	4.51
	ML	0.7344	6.02

Comparative Genomics

Alignments between *Brachypodium* v1.0 genes, and the genes predicted in the build 5 rice pseudomolecules (www.tigr.org) and 10 sorghum pseudomolecules (www.phytozome.net) were generated. A set of 6,426 wheat ESTs representing 15,569 loci mapped to Chinese Spring deletion bins [61] were downloaded from the GrainGenes website (http://wheat.pw.usda.gov/). The Triticeae comparative mapping set comprised a set of 5,003 curated non-redundant ESTs generated from these [62], and genetic maps of 1,015 barley ESTs [63] and 863 *Ae. tauschii* ESTs [64]. Gene relationships and order were compared using the CIP-CALP method [62]. Syntenic blocks were defined precisely between 25,532 annotated *Brachypodium* protein-coding genes, 7,216 sorghum orthologs (12 syntenic blocks), 8,533 rice orthologs (12 syntenic blocks) and 2,516 Triticeae orthologs (12 syntenic blocks).


56-73 MYA
n=5
A5 A7 A11 A8 A4
n=10
A1 A5 A7 A10 A11 A12 A8 A9 A4 A6
Paleotetraploidy
n=12
A1 A5 A7 A10 A11 A12 A8 A9 A4 A6
45-60 MYA
40-54 MYA
32-39 MYA
Ehrhartoideae
Intermediate n=12
A1 A5 A7 A10 A3 A11 A12 A8 A9 A4 A6 A2
Brachypodium
Intermediate n=5
A1 A5 A7 A10 A3 A11 A12 A8 A9 A4 A6 A2
A1 A5 A3 A6 A7 A11 A12 A9 A10 A8 A2 A4
Poideae
Intermediate n=7
A1 A5 A7 A10 A3 A11 A12 A8 A9 A4 A6 A2
A1 A10 A5 A3 A11 A12 A9 A7 A4 A6 A8 A2
Panicoideae
Intermediate n=10
A1 A5 A7 A10 A3 A11 A12 A8 A9 A4 A6 A2
A3 A10 A9 A7
12 MYA
5 MYA
Maize
Intermédiaire n=20
A6 A3-10 A8 A6 A5 A1 A5 A1 A1 A12 A1 A11 A8 A2
A3-10 A6 A5 A3-10 A8 A12 A3-10 A12 A8 A4 A4 A11 A12 A7-9 A7-9 A7-9
Rice
r1 r5 r7 r10 r3 r11 r12 r8 r9 r4 r6 r2
Brachypodium
bd1 bd2 bd3 bd4 bd5
Triticeae
t1 t2 t3 t4 t5 t6 t7
Sorghum
s1 s2 s3 s4 s5 s6 s7 s8 s9 s10
Maize
m1 m2 m3 m4 m5 m6 m7 m8 m9 m10

Supplementary Figure 18. Grass chromosome evolution model. The monocot chromosomes (r1-r12 for rice, t1-t7 for Triticeae, bd1-bd5 for *Brachypodium,* s1-s10 for sorghum, and m1-m10 for maize) are represented with a five colour code to illustrate the evolution of segments from a common ancestor with five proto-chromosomes and a n=12 intermediate as described in [62], and are named according to the rice nomenclature. The events that have shaped the structure of the 5 different grass genomes including the 7 *Brachypodium* chromosome nested insertion events during their evolution from the common ancestor are indicated as whole genome duplication, ancestral chromosome translocations and fusions, and lineage- specific nested chromosome insertions.

Supplementary Table 17. Accelerated genome evolution in the pooid grasses. Numbers and rates per million years of inversions and subchromosomal size translocations and all structural changes (including chromosome size translocations) detected in comparisons of the *Ae. tauschii* genetic map with the sorghum, rice and *Brachypodium* genome sequences.

Internode	Time* (MY)	Inversions and subchrom. translocations (No.)	Rate No. changes MY^{-1}	All changes (No.)	Rate No. changes MY^{-1}
Brachypodium	35.8	5	0.14	12	0.34
Ae. tauschii	35.8	36	1.01	41	1.15
Brachypodium + *Ae. tauschii*	11.5	1	0.09	1	0.09
Rice	47.3	4	0.08	4	0.08
Sorghum	52.7	5	0.09	7	0.13
Could not be assigned		7		7	

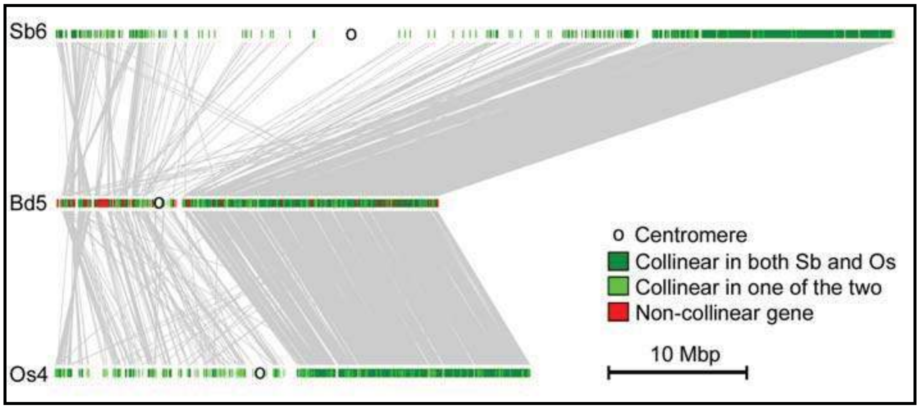
*Divergence times are an average of the times calculated by the NG and ML methods (Supplementary Table 16).

The linear order of 863 gene loci mapped on the *Ae. tauschii* EST genetic map [64] and orthologous loci in *Brachypodium*, rice and sorghum were used to estimate the rates of chromosome evolution at the internodes of their phylogenetic tree (Figure 3A). The following strategy was used to assign changes in gene collinearity due to inversions and translocations into the tree internodes. If gene order in a single genome differed from the remaining three, the structural change was assigned to the appropriate terminal internode. If gene order was collinear in the *Ae. tauschii* and *Brachypodium* genomes, but differed from that in rice and sorghum, the change was assigned to the internal internode in the tree between the divergence of *Ae. tauschii* and *Brachypodium* on one side and the divergence of Pooideae (*Brachypodium* + *Ae. tauschii*) and Ehrhartoideae (rice) on the other side. No structural change was found in *Ae. tauschii* or *Brachypodium* that was shared with sorghum but was absent from rice, consistent with the phylogenetic tree in Figure 3A. Due to the absence of an outgroup, it was not possible to discriminate between structural changes that took place after the divergence of sorghum from the common ancestor of *Ae. tauschii*, *Brachypodium* and rice and before the divergence of rice from the ancestor of *Brachypodium* and *Ae. tauschii*, and those that took place in the sorghum branch; all such changes were assigned to the sorghum terminal branch. The rate of chromosome evolution in the sorghum lineage may therefore be slightly inflated. A total of 51 inversions and subchromosomal-size translocations could be assigned to internodes of the phylogenetic tree; seven small inversions could not be assigned because of the lack of recombination between relevant markers in the *Ae. tauschii* mapping population. In addition to the sub-chromosome sized changes, 14 chromosome-size translocations resulting in the dysploid reductions of the basic chromosome number were assigned to three terminal internodes (Supplementary Table 17). It was assumed in the computation of the chromosome evolution rates that the number of genes in a genome that could be subjected to a structural change has remained more-or-less constant during the phylogeny of the four genomes. A linear relationship was therefore assumed between the accumulation of structural changes in an internode of the tree and time, and the rate of chromosome evolution per million years (MY) was computed by dividing the number of structural changes in a specific internode by the internode length in MY.

Supplementary Table 18. Large *Brachypodium* gene families and their degree of collinearity in rice and sorghum. The location of members of seven large gene families were compared to determine if the degree of collinearity correlates with the degree of sequence conservation. Note that the highly variable F-box and NBS-LRR gene families also have the least conservation of collinearity.

Gene family	total	collinear in one[1]	collinear in both[2]
HSP40	106	90.6%	76.4%
RINGFYVEHPD	384	89.8%	69.8%
Ser/Thr kinase	904	83.5%	64.2%
WD40YVTN	160	81.9%	61.9%
Cytochrome P450	261	66.7%	45.2%
F-box	301	57.1%	20.6%
NBS-LRR	178	52.7%	12.6%

[1] Percentage of genes found in collinear position in either rice or sorghum.
[2] Percentage of genes found in collinear position in both rice and sorghum.



Supplementary Figure 19. Map of *Brachypodium* chromosome 5 (Bd5) and its syntenic chromosomes from sorghum (Sb6) and rice (Os4). Collinear genes are connected by grey lines. In all three species the short arm has lower gene density, reduced collinearity and multiple rearrangements such as inversions and translocations. The short arm of Bd5 has the lowest ratio of intact:solo LTR elements (0.89 vs 2.6 for the whole genome), indicating a gain of retrolements.

Acknowledgements

Sequencing was supported by the US Department of Energy Joint Genome Institute Community Sequencing Program Contract 776757 to J.P.V, D.F.G., T.C.M. and M.W.B. Work at the John Innes Centre was supported by BBSRC grant BB/E004725/1,and EU Contracts Agronomics (FP6-037704) and TriticeaeGenome (FP7-212019) to M.W.B., and BBSRC Research Studentship BBS/S/E/2006/13204 to J.W. Work conducted at the WRRC was supported by USDA-ARS CRIS projects 532502100-010, 533502100-011, 532502100-013 and by NSF grant DBI-0638558. Work at MIPS was supported by funds from EU TriticeaeGenome (FP7-212019) and GABI Barlex funded by the German Ministry of Education and Research (BMBF FKZ 0314000). Work at MTT Agrifood Research Finland and at the University of Helsinki was supported in part by the EU TriticeaeGenome project (FP7-212019). Illumina transcriptome sequencing was supported by DOE Plant Feedstock Genomics for Bioenergy grant DE-FG02-08ER64630 and Oregon State Agricultural Research Foundation grant ARF4435 to T.C.M., and a Computational and Genome Biology Initiative Fellowship from Oregon State University to H.D.P. The small RNA research was supported by DOE Plant Feedstock Genomics for Bioenergy grants DE-FG02-07-ER64450 and DE-FG02-08ER64630 to P.J.G. and T.C.M, respectively, and NSF grants MCB-0618433 and PGRP-0701745 to J.C.C. and B.C.M., respectively. Secretome analysis was supported by NSF Plant Genome Research Program grant DBI-0606595 to J.K.C.R. Manual annotation was supported by USDA NRI Grant 2008-35600-18783 to J.H.C., Office of Science (BER), U. S. Department of Energy, Interagency Agreement No. DE-AI02-07ER64452 to J.P.V.; a Villum-Kann Rasmussen grant to the Pro-Active Plant Centre and by the U. S. Department of Energy, Office of Science, through contract DE-AC02-05CH11231 to H.V.S; and NNSF grant 30425039,30771341, State Transgenic Project 2009ZX08009-048B, and Introducing Talents of Discipline Program to Universities grant 111-2-03 to Z.L. Work at INRA was supported by French ANR-KBBE Grant Cellwall (ANR 08-KBBE-04) and EU Contracts Agronomics (FP6-037704) and Renewall (FP7-211981) to H.H.

We thank Mayumi Nakano and Caghan Demirci (Delaware Biotechnology Institute, University of Delaware) for parsing the genome files for the small RNA analysis. We thank Marianne Smith, Lisa Coffey, and Cheng-Ting "Eddy" Yeh (Center for Plant Genomics, Iowa State University) for technical assistance with 454 sequencing. We thank Mark Dasenko and Steve Drake (Center for Genome Research and Biocomputing, Oregon State University) for assistance with Illumina sequencing and bioinformatics. We thank David Goodstein and Shengqing Shu (JGI) for support of www.phytozome.net/brachy. Finally, we thank Leonid Brodsky (University of Haifa) for calculating the LTR retrotransposon half-lives.

References

1. Peterson, D.G., Boehm, K.S. & Stack, S.M. Isolation of milligram quantities of DNA from tomato (Lycopersicon esculentum), a plant containing high levels of polyphenolic compounds. *Plant Molecular Biology Reporter* **15**, 148-153 (1997).
2. Gu, Y.Q. et al. A BAC-based physical map of *Brachypodium* distachyon and its comparative analysis with rice and wheat. *BMC Genomics* **10**, 496 (2009).
3. Huo, N. et al. Construction and characterization of two BAC libraries from *Brachypodium* distachyon, a new model for grass genomics. *Genome* **49**, 1099-108 (2006).
4. Huo, N. et al. The nuclear genome of *Brachypodium* distachyon: analysis of BAC end sequences. *Funct Integr Genomics* **8**, 135-47 (2008).
5. Jaffe, D.B. et al. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res* **13**, 91-6 (2003).

6. Hasterok, R. et al. Alignment of the genomes of *Brachypodium* distachyon and temperate cereals and grasses using bacterial artificial chromosome landing with fluorescence in situ hybridization. *Genetics* **173**, 349-62 (2006).
7. Jenkins, G. & Hasterok, R. BAC 'landing' on chromosomes of *Brachypodium* distachyon for comparative genome alignment. *Nat Protoc* **2**, 88-98 (2007).
8. Unfried, I. & Gruendler, P. Nucleotide sequence of the 5.8S and 25S rRNA genes and of the internal transcribed spacers from Arabidopsis thaliana. *Nucleic Acids Res* **18**, 4011 (1990).
9. Gerlach, W.L. & Dyer, T.A. Sequence organization of the repeating units in the nucleus of wheat which contain 5S rRNA genes. *Nucleic Acids Res* **8**, 4851-65 (1980).
10. Fox, S., Filichkin, S. & Mockler, T. (eds.). *Applications of ultra high throughput sequencing*, (Humana Press, 2009).
11. Llave, C., Kasschau, K.D., Rector, M.A. & Carrington, J.C. Endogenous and silencing-associated small RNAs in plants. *Plant Cell* **14**, 1605-19 (2002).
12. Kasschau, K.D. et al. Genome-wide profiling and analysis of Arabidopsis siRNAs. *PLoS Biol* **5**, e57 (2007).
13. Lu, C. et al. Elucidation of the small RNA component of the transcriptome. *Science* **309**, 1567-9 (2005).
14. Lu, C., Meyers, B.C. & Green, P.J. Construction of small RNA cDNA libraries for deep sequencing. *Methods* **43**, 110-7 (2007).
15. Howell, M.D. et al. Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in Arabidopsis reveals dependency on miRNA- and tasiRNA-directed targeting. *Plant Cell* **19**, 926-42 (2007).
16. Lister, R. et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**, 523-36 (2008).
17. Gremme, G., Brendel, V., Sparkes, M. & Kurtz, S. Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology* **47**, 965-978 (2005).
18. Ouyang, S. et al. The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res* **35**, D883-7 (2007).
19. Tanaka, T. et al. The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res* **36**, D1028-33 (2008).
20. Paterson, A.H. et al. The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551-6 (2009).
21. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815 (2000).
22. Rhee, S.Y. et al. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res* **31**, 224-8 (2003).
23. Allen, J.E. & Salzberg, S.L. JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics* **21**, 3596-603 (2005).
24. Mayer, K. et al. Sequence and analysis of chromosome 4 of the plant Arabidopsis thaliana. *Nature* **402**, 769-77 (1999).
25. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-64 (1997).
26. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-9 (2006).
27. Gotz, S. et al. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res* **36**, 3420-35 (2008).
28. Fincher, G.B. Exploring the evolution of (1,3;1,4)-beta-D-glucans in plant cell walls: comparative genomics can help! *Curr Opin Plant Biol* **12**, 140-7 (2009).
29. Sarria, R. et al. Characterization of a family of Arabidopsis genes related to xyloglucan fucosyltransferase1. *Plant Physiol* **127**, 1595-606 (2001).

30. Mitchell, R.A., Dupree, P. & Shewry, P.R. A novel bioinformatics approach identifies candidate genes for the synthesis and feruloylation of arabinoxylan. *Plant Physiol* **144**, 43-53 (2007).
31. Imaizumi, T. & Kay, S.A. Photoperiodic control of flowering: not only by coincidence. *Trends Plant Sci* **11**, 550-8 (2006).
32. Colasanti, J. & Coneva, V. Mechanisms of floral induction in grasses: something borrowed, something new. *Plant Physiol* **149**, 56-62 (2009).
33. Dardick, C., Chen, J., Richter, T., Ouyang, S. & Ronald, P. The rice kinase database. A phylogenomic database for the rice kinome. *Plant Physiol* **143**, 579-86 (2007).
34. Dardick, C. & Ronald, P. Plant and animal pathogen recognition receptors signal through non-RD kinases. *PLoS Pathog* **2**, e2 (2006).
35. Shiu, S.H. et al. Comparative analysis of the receptor-like kinase family in Arabidopsis and rice. *Plant Cell* **16**, 1220-34 (2004).
36. Escobar-Restrepo, J.M. et al. The FERONIA receptor-like kinase mediates male-female interactions during pollen tube reception. *Science* **317**, 656-60 (2007).
37. Hematy, K. & Hofte, H. Novel receptor kinases involved in growth regulation. *Curr Opin Plant Biol* **11**, 321-8 (2008).
38. Vogel, J. Unique aspects of the grass cell wall. *Curr Opin Plant Biol* **11**, 301-7 (2008).
39. Enright, A.J., Kunin, V. & Ouzounis, C.A. Protein families and TRIBES in genome sequence space. *Nucleic Acids Res* **31**, 4632-8 (2003).
40. Cosgrove, D.J. Loosening of plant cell walls by expansins. *Nature* **407**, 321-6 (2000).
41. Cantarel, B.L. et al. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res* **37**, D233-8 (2009).
42. McCarthy, E.M. & McDonald, J.F. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* **19**, 362-7 (2003).
43. Ma, J. & Bennetzen, J.L. Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci U S A* **101**, 12404-10 (2004).
44. Wicker, T. & Keller, B. Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. *Genome Res* **17**, 1072-81 (2007).
45. Wicker, T. et al. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**, 973-82 (2007).
46. Maia, L.d.C. et al. SSR Locator: Tool for Simple Sequence Repeat Discovery Integrated with Primer Design and PCR Simulation. *Int J Plant Genomics* **2008**, 412696 (2008).
47. Varshney, R.K., Thiel, T., Stein, N., Langridge, P. & Graner, A. In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cell Mol Biol Lett* **7**, 537-46 (2002).
48. Temnykh, S. et al. Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res* **11**, 1441-52 (2001).
49. Field, D. & Wills, C. Long, polymorphic microsatellites in simple organisms. *Proc Biol Sci* **263**, 209-15 (1996).
50. Jurka, J. & Pethiyagoda, C. Simple repetitive DNA sequences from primates: compilation and analysis. *J Mol Evol* **40**, 120-6 (1995).
51. Li, L., Stoeckert, C.J., Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-89 (2003).
52. Tatusova, T.A. & Madden, T.L. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett* **174**, 247-50 (1999).
53. Kaplinsky, N.J., Braun, D.M., Penterman, J., Goff, S.A. & Freeling, M. Utility and distribution of conserved noncoding sequences in the grasses. *Proc Natl Acad Sci U S A* **99**, 6147-51 (2002).

54. Lyons, E. & Freeling, M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J* **53**, 661-73 (2008).

55. Liu, Q. et al. Two transcription factors, DREB1 and DREB2, with an EREBP/AP2 DNA binding domain separate two cellular signal transduction pathways in drought- and low-temperature-responsive gene expression, respectively, in Arabidopsis. *Plant Cell* **10**, 1391-406 (1998).

56. Degenkolbe, T. et al. Expression profiling of rice cultivars differing in their tolerance to long-term drought stress. *Plant Mol Biol* **69**, 133-53 (2009).

57. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276-7 (2000).

58. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-91 (2007).

59. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**, 418-26 (1986).

60. Wolfe, K.H., Sharpe, P.M. & Li, W.H. Rates of synonymous substitutions in plant nuclear genes. *Journal of Molecular Evolution* **29**, 208-211 (1989).

61. Qi, L.L. et al. A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. *Genetics* **168**, 701-12 (2004).

62. Salse, J. et al. Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* **20**, 11-24 (2008).

63. Stein, N. et al. A 1,000-loci transcript map of the barley genome: new anchoring points for integrative grass genomics. *Theor Appl Genet* **114**, 823-39 (2007).

64. Luo, M.C. et al. Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in Triticeae. *Proc Natl Acad Sci U S A* **106**, 15780-5 (2009).