Genome sequencing and analysis of the model grass Brachypodium distachyon

August 9, 2009

Correspondence and requests for materials should be addressed to one or all of the following:

John P. Vogel - Email: john.vogel@ars.usda.gov
David F. Garvin - Email: david.garvin@ars.usda.gov
Todd C. Mockler - Email: tmockler@cgrb.oregonstate.edu
Michael W. Bevan - Email: michael.bevan@bbsrc.ac.uk

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 (Leader)[4], 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 (Leader)[1], 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 (Leader)[3], 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 (Leader)[16], 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 (Leader)[20], Jan P. Buchmann[20], Jaakko Tanskanen[21], Alan H. Schulman (Leader)[21], 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 (Leader)[27], 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 (Leader)[30]

Genome sequencing and analysis of the model grass Brachypodium distachyon

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 (Leader)[1], 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[8,31,46], 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. Hudson-Alpha Institute, Huntsville, Alabama 35806, USA.
5. US DOE Joint Genome Institute/Lawrence Berkeley National Lab, 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.
46. The School of Plant Molecular Systems Biotechnology, Kyung Hee University, Yongin 446-701, Korea.

Acknowledgement

The work conducted by the US Department of Energy (DOE) Joint Genome Institute is supported by the Office of Science of the DOE under Contract Number DE-AC02-05CH11231. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government, or any agency thereof, or the Regents of the University of California.

This work was supported by the US Department of Energy Joint Genome Institute Community Sequencing Program Contract 776757 to J.D.V, D.F.G., T.C.M. and M.W.B., BBSRC grant BB/E004725/1 to M.W.B. and BBSRC Research Studentship BBS/S/E/2006/13204 to J.W. Work at MIPS was supported by funds from EU Triticeae Genome (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 Triticeae Genome project (FP7-212019). Illumina transcriptome sequencing was supported by DOE Plant Feedstock Genomics for Bioenergy grant DE-FG02-08ER64630 and Oregon State Agricultural

Genome sequencing and analysis of the model grass Brachypodium distachyon

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. Manual annotation was supported by USDA NRI Grant 2008-35600-18783 to J.H.C., 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. 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.

DISCLAIMER:

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

version v9 aug 7

Genome sequence analysis of the model grass *Brachypodium distachyon*: insights into grass genome evolution.

The International Brachypodium Initiative*

**A list of participants and affiliations appear at the end of the paper*

Abstract

Three subfamilies of grasses, the Erhardtoideae (rice), the Panicoideae (maize, sorghum, sugar cane and millet), and the Pooideae (wheat, barley and cool season forage grasses) provide the basis of human nutrition and are poised to become major sources of renewable energy. Here we describe the complete genome sequence of the wild grass *Brachypodium distachyon* (Brachypodium), the first member of the Pooideae subfamily to be completely sequenced. Comparison of the Brachypodium, rice and sorghum genomes reveals a precise sequence- based history of genome evolution across a broad diversity of the grass family and identifies nested insertions of whole chromosomes into centromeric regions as a predominant mechanism driving chromosome evolution in the grasses. The relatively compact genome of Brachypodium is maintained by a balance of retroelement replication and loss. The complete genome sequence of Brachypodium, coupled to its exceptional promise as a model system for grass research, will support the development of new energy and food crops. *155 words*

Introduction

The "rapid rise and early diversification" of flowering plants approximately 90-130 MYA (Million Years Ago) (*1*) was said by Darwin to be an "abominable mystery"(*2*). The grass family (Poaceae) exemplifies this extreme diversification, having evolved from a common ancestor between 55-70 MY ago to form 600-700 diverse genera and over 10,000 species that today dominate many different ecological and agricultural systems (*3*). Several grass species have

been domesticated during human history to provide the bulk of human and animal nutrition. Furthermore, because of their very high productivity and adaptability, grass crops are also promising sources of sustainable energy (*4*). This has led to intense research aimed at improving grass crops for sustainable grain, forage and energy production.

Three diverse subfamilies of grasses currently provide our main food and feed sources: the Erhardtoideae (rice); the Panicoideae (maize, sorghum, sugar cane and millet); and the Pooideae (wheat, barley and cool season forage grasses). To date the rice (*5*) and sorghum (*6*) genomes have been completely sequenced and analyzed. Comparison of these genomes and the physical map of maize identified an ancestral whole genome duplication (*7*) and post-duplication gene loss. Nevertheless, an extensive conservation of gene order is maintained (*6, 8*). Despite these analyses, the mechanisms shaping grass genomes remain poorly understood.

Most of the cool season cereal, forage and turf grasses belong to the subfamily Pooideae, which is also the most diverse grass subfamily with over 3,000 species. The genomes of some pooids are characterized by extreme size and complexity; for example, the hexaploid bread wheat genome is approximately 17,000 Mb and contains three independent genomes. Such large and complex genomes have prohibited genome-scale comparisons spanning the three most important grass subfamilies; consequently it has not been possible to identify systematically gene functions in this important grass subfamily using genomic methods.

Here we describe the complete genome sequence of *Brachypodium distachyon* (Brachypodium or purple false brome), the first member of the Pooideae subfamily and the first wild grass species to be completely sequenced. Brachypodium is an annual grass endemic to the Mediterranean and Middle East (*9*). It is an exceptionally promising model system for the grasses because it possesses many of the attributes (rapid life cycle, simple growth requirements, small stature, self fertility, small genome and highly efficient transformation (*10, 11*) that have made Arabidopsis a powerful model species for dicots. These features contributed to the promotion of Brachypodium as a model for the grasses (*9, 12*) The Brachypodium genome sequence described here permits for the first time whole genome comparisons between members of the three most economically important grass subfamilies, represents a

major advance for grass functional genomics, and provides a template for analysis of the large and complex genomes of other pooid grasses. *442 words*

Assembly of chromosome- scale features from whole genome shotgun sequence.

Diploid inbred line Bd21(*12*) derived from USDA accession PI 254867 collected near Mosul in northern Iraq was sequenced. A whole genome shotgun (WGS) sequencing strategy utilized three sized subclone libraries in addition to BAC-end sequence (BES) generated from four different libraries (Table S1). This produced an initial assembly of 1,754 contigs in 83 scaffolds using Arachne (*13*). Remarkably, the 10 largest scaffolds contained 99.6% of the sequence (Table S2.). Alignment with 562 markers on a genetic map (Figure S1) detected two false joins and created an additional seven joins. The final assembly covered 271 Mb to a final depth of 9.4x with only 0.4% gaps (Table S3). This size falls within the range of diploid Brachypodium genome sizes measured by flow cytometry (*14, 15*). The sequence assembly was confirmed by alignment with BAC end sequences from two physical map of BACs and cytogenetic analysis using physically- mapped BACs as FISH (fluorescent in situ hybridization) probes further confirmed these assemblies (*16*) (Figure S2). The arms of chromosomes 1 to 4 were covered by 11 scaffolds containing proximal centromeric repeats and distal subtelomeric repeats (Figure 1). Over 95% of ESTs and transcript consensus sequences were mapped to the 11 genome sequence scaffolds, indicating high coverage (Figure S3). The shortest chromosome, 5, was covered by a single 28 Mb scaffold containing a central array of typical centromeric satellite repeats and terminating in 25S ribosomal repeats on the short arm and subtelomeric repeats on the long arm. Compared to other grasses, the Brachypodium genome has a very compact structure, with retrotransposons concentrated at the centromeres and syntenic breakpoints (Figures 1 and 5), with extensive regions of high gene density towards the telomeres, and a broad distribution of DNA transposons and derivatives that are primarily associated with gene-rich regions.

Gene annotation and analysis

A total of 25,532 protein coding gene loci was predicted in the v1.0 annotation of the Brachypodium genome as described in (*17*) (Table 1). This is in the same range as rice (RAP2, 28,236) (*18*) and sorghum (v1.4, 27,640) genes (*6*), indicating a haploid protein- coding gene content between 25,000- 30,000 genes across the broad diversity of grasses. Gene predictions were supported by protein and transcript databases and *ab initio* gene finders. Brachypodium ESTs from six different tissues and multiple growth conditions were generated by 454 and Sanger sequencing methods (Table S4). This evidence was incorporated into a statistical combiner trained on a manually curated set of genes and applied to the complete genome sequence to derive a unified gene model from weighted initial predictions for each locus. Coding structures were subsequently post-processed with EST data to fit models closely to transcript evidence. A total of 32,255 transcript models including splice variants was identified in Brachypodium. The gene models were evaluated for transcript support using ~10.2 Gb Illumina sequence generated using RNA-seq (*19*). Overall, 92.7% of predicted CDS (coding sequences) were supported by Illumina reads matching two or more unique locations within the predicted CDS, and the median coverage over the lengths of the predicted CDS was 91% (Figure 2A; Figure S3). The extensive experimental support provided by Illumina transcriptome sequence underscores the exceptionally accurate set of Brachypodium gene predictions. These can be browsed and downloaded from several genome databases (*20*).

We validated and improved gene predictions by manually annotating 2,755 gene models from 72 diverse gene families using multiple transcript sequences and alignment to genes from other organisms (Table S5). Only 13% of the gene models examined were modified, demonstrating the accuracy of the automated gene predictions. We emphasized gene families relevant to bioenergy research (*4*), included genes involved in the biosynthesis and remodeling of the cell wall (cellulose synthase (CS, 10 genes), cellulose synthase-like (CSL, 25 genes), other glycosyltransferases (GT, 313 genes), glycosyl hydrolases (GH, 339 genes), and 179 genes putatively involved in monolignol or pectin metabolism). We identified and annotated 802 transcription factors from 16 families according to community standards (*21*). Phylogenetic trees for 62 gene families were constructed using genes from rice, Arabidopsis, sorghum and poplar.

In nearly all cases, Brachypodium had a similar distribution of gene family members within the trees as rice and sorghum, demonstrating the essential unity of grass genomes. Some differences were identified (Figures S5 and S6); CSL subfamily J was proposed as a clade specific to some grasses including maize, sorghum, barley, and wheat, but not rice, Brachypodium or dicots (*22*). However, our analysis revealed that Brachypodium, poplar and several other dicots had CSLJ genes (Figure S6). Using BLAST scores and pfam domains, we placed a further 2,755 gene models in 13 gene families including kinases, proteasome subunits, auxin signaling genes and F-box proteins (Table S6). 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 (Table S7). Using domain scans of unmasked genome sequence we identified an additional 170 putative F-box containing genes and 67 putative BTB genes and brought these gene family numbers into a broad agreement with other plants.

We compared the predicted secreted proteomes of Brachypodium, Arabidopsis, and rice to examine whether the substantial differences between grass and dicot cell walls (*23*) are correlated with distinctive populations of secreted proteins. There were significant differences between Arabidopsis and the grasses, mirroring the differences in cell wall architecture (Tables S8, S9 and Figure S7). Furthermore, signal peptide probability curves of the predicted proteomes of Brachypodium and Arabidopsis were more similar to each other than to rice, suggesting accurate prediction of Brachypodium start codons (Figure 2B).

The complete gene sets of rice, sorghum, and Brachypodium and multiple ESTs from wheat and barley were compared using OrthoMCL to identify pooid-specific gene families (*17*). Figure 2C shows that between 77-84% of gene families are shared among all three grass groups, reflecting their relatively recent common origin. We identified core- and lineage- enriched gene clusters (Figure S8) that were assigned molecular functions using the blast2go suite (Table S10) and Pfam domains. The broad biological functions of a monocot core gene set were distinguished from an angiosperm core set by an over-representation of transmembrane receptor-like kinases, secondary metabolism enzymes, transcription factors and sugar

transferases. This reflects the specific secondary metabolism, defense, development and cell wall synthesis pathways of the grasses. The pooid-specific core was enriched for heme-binding proteins, receptor kinases, ion- and cation- binding proteins and glycosyl transferases. These are involved in secondary metabolite production, cell wall formation and possible adaptation to soils. Brachypodium- enriched gene functions also included P450 proteins and defense-related enzyme activities such as peroxidases and peptidases, and adaptive functions such as metal binding. The gene classes enriched in the monocot core set had a highly significant increased proportion of tandem genes, demonstrating a prominent role for tandem gene expansion in the evolution of monocot-specific genes (Figure S9 and Table S11).

The compact genome of Brachypodium is maintained by balancing retroelement replication and loss.

The replicative life cycle of active retrotransposon families can lead to increased copy number and genome expansion in grasses (*24*). To understand the basis of the relatively compact genome of Brachypodium, we conducted an exhaustive analysis of all transposable element classes in the genome (*17*)(Table 2). 690 intact LTR retrotransposons occupy 6.50 Mbp (2.4%) of the Brachypodium genome (Table 2), compared with 2.8% in Arabidopsis. These solo LTRs and other retrotransposon fragments comprise 21.4% of the genome, 26% in rice, 54% in Sorghum, and over 80% in wheat. *Gypsy* and *Copia* solo LTRs have similar relative abundance and are on average 4.3 MY old, similar to the ~3 MY persistence time in rice (*25*). Thirteen retroelement clusters were younger than 20,000 years, showing an abrupt recent activation compared to rice (*26*) (Figure 3A), and a further 53 retroelement clusters were less than 0.1 MY old. Two of the most recently active *Angela–BARE–Wis* family elements have multiple solo LTRs associated with them, demonstrating active retrotransposon loss through recombination (*27*). A minimum of 17.4 Mb (nearly 30% of the repeat content and 6% of the genome) has been lost by LTR:LTR recombination, demonstrating that active retroelement expansion is countered by efficient removal by recombination to maintain a compact genome. In contrast, a similar assessment of retroelements in the Triticeae indicated a very long persistence, too long to be calculated from the available dataset (*26*).

Class 2 transposons reveal a multitude of complex interaction between autonomous and non-autonomous elements

We identified a total of 29,630 DNA transposons (Class 2 repeats) belonging to 253 families and 6 super-families (Table 2) comprising 4.77% of the *Brachypodium* genome sequence, comparable to the 2.7% to 13.7% that are found in other grass genomes (*6, 28*). In the *Mariner DTT* superfamily we identified 52 potentially autonomous "Mother" elements and 20,994 derivative *Stowaway* MITEs (*29*) (21 families, 0.88% of the genome). In contrast, the *Harbinger* superfamily contains a Mother population of 862 elements and only 2,569 derivative *Tourist* MITES (*30*) (19 families, 0.18% of the genome). Many apparently non-autonomous elements appear to recruit enzymes for transposition across family boundaries to function as semi-autonomous Mother elements (Figure 3B). Only three *Mariner* and six *Harbinger* Mother elements had matches to Brachypodium transcriptome data, indicating that the proliferation of many thousands of non–autonomous elements depends on a few functional Mother elements. This is similar in rice and sorghum, although analysis of these two genomes was less exhaustive (*6, 31*). We conclude that grass genomes can only tolerate a very small number of active Mother elements because of the potentially disruptive effects of MITEs on the genome. In other grasses *Mutator*, *Helitron* and *CACTA* transposons are responsible for at least partial replication of hundreds of genes (*32, 33*). Brachypodium *Helitrons* are less numerous than in other grasses and were not found to carry gene fragments. In contrast, two *CACTA DTC* families (M and N) were found to carry a total of 5 non-element genes. The *Harbinger U* family has amplified a particular NBS-LRR gene family with which it has undergone a gene fusion (Figure 3B); EST data shows this is also present in wheat and barley. This adds *Harbingers* to the group of transposable elements implicated in gene mobility.

Conserved non-coding sequences and simple sequence repeats

Conserved noncoding sequences (CNSs) are a subclass of phylogenetic footprints conserved during evolution (*34*). Possible functional roles of CNSs include interactions with transcription factors. We identified 18,664 sequence regions that are conserved between orthologous genes in Brachypodium, sorghum and rice. These were classified as "true CNSs" (11,328 sequences),

where the conserved sequences are syntenic in the three genomes, or “simple conserved regions” (7,336 sequences) where regions are conserved within the gene space of orthologs but are not syntenic (Figures S10-S12). We identified potentially functional elements within these CNSs such as GCCGAC elements, previously shown to bind DREB transcription factors that activate drought responses. The Brachypodium genome contains a total of 98,027 loci of simple sequence repeats (monomers to hexamers) comprising a total of 1.4 Mb or 0.51 % of the genome (Table S12), an average of 359 SSRs/Mb. By comparison, Arabidopsis and rice show at least twice this abundance, 755 and 686 SSRs/Mb respectively.

Whole genome sequence-level comparison across the diversity of grass genomes

Brachypodium is the first pooid grass to be sequenced, enabling comparison of genome features across three of the major grass subfamilies. The evolutionary relationships between Brachypodium, sorghum, rice and wheat were assessed by measuring the mean synonymous substitution rates (Ks) of orthologous gene pairs (*17*) (Figure S13 and Table S13). The distribution maxima provide estimates of divergence times of Brachypodium from wheat 29.4 (±4.9) MYA Million years ago), rice 42.1 (±6.9) MYA and sorghum 50.5 (±7.5) MYA (Figure 4A). The distribution of synonymous rates of orthologous gene pairs in the intra-genomic Brachypodium duplications (Figure 4B) suggests duplication ~65-73 MYA ago, prior to the diversification of the grasses. This is consistent with previous evolutionary histories inferred from relatively small numbers of chloroplast and nuclear genes (*35-37*) and the *Hardness* locus (*38*), and provides a more precise range of divergence times.

Using the rice and sorghum genome sequences, genetically mapped barley (*39*) and *Aegilops tauschii* (the D genome donor of hexaploid wheat) ESTs (*40*), bin-mapped wheat ESTs (*41*) and robust alignment criteria (*42*), we identified 21,045 orthologous relationships between Brachypodium / rice/ sorghum / Triticeae and 723 paralogous relationships among Brachypodium chromosomes (*17*). The paralogous relationships revealed six major inter-chromosomal duplications covering 99.7% of the genome (Figure 4B) that represent ancestral whole genome duplication (*43*). The orthologous relationships indentified 59 blocks of collinear

genes covering 99.42% of the Brachypodium genome (Figures 4C, E and F). These relationships are consistent with an evolutionary scenario that shaped five Brachypodium chromosomes from a five chromosome grass ancestral genome *via* a 12 chromosome intermediate involving seven major chromosome fusions (*42*). These collinear gene blocks provide a robust and precise sequence framework for understanding genome evolution across a broad diversity of economically important grasses, for identifying candidate genes and for interpreting genome sequence assemblies from other pooid grasses.

We identified 14 major syntenic disruptions between Brachypodium and rice/sorghum that can be explained by seven precisely nested fusions of entire chromosomes into centromeric regions (Figure 5) (*3*). Figure 4D illustrates how the order of collinear gene blocks also supports this interpretation. Brachypodium chromosomes 1, 3 and 4 are the product of two nested fusions each and the structure of chromosome 2 can be explained by a single fusion event. In contrast chromosome 5 has remained intact during the evolution of the grasses. We identified similar nested insertions in sorghum chromosomes 1 and 2, and, using genetic mapping data, identified nested insertions in barley chromosomes 1, 2 and 4 (Figure 5C and 5D). This explains retroelement distribution patterns on chromosomes 1-4 at the boundaries of insertions that preserve higher gene density at the former distal regions of the inserted chromosome (Figure 1). Our analysis suggests that nested fusions are the predominant mechanism of chromosome fusion in grasses, in contrast to dicots where chromosome fusions occur most often at chromosome ends (*44*). Brachypodium gene order was compared with 12 sequenced syntenic regions of wheat and barley covering a total of 1.9 Mb (Figure S14); this revealed 62.5% conservation of gene order. A similar comparison to rice and sorghum revealed 55% conservation, consistent with the closer evolutionary relationships of Brachypodium to wheat and barley. This illustrates the potential for Brachypodium sequence to aid gene discovery in other pooid grasses.

Interestingly, comparison of evolutionary rates between Brachypodium, sorghum, rice, and *Ae. tauschii* demonstrated a substantially higher rate of genome change in *Ae. tauschii* (Table S14). This could be due to retroelement activity that increases the rate of syntenic disruption in larger

genomes, such as we propose on chromosome 5S below (*45*). Several large gene families exhibit different extents of conserved gene order in Brachypodium, sorghum and rice. Table S15 shows that among seven relatively large gene families, four exhibit a high conservation of gene order; in contrast NBS-LRR disease resistance genes and F-box gene family members examined were almost never found in syntenic order. This is consistent with the rapid diversification of these gene families by recombination (*46*).

Chromosome 5 is a bad neighborhood for genes

The short arm of chromosome 5 (Bd5S) has a gene density of only 53.7 genes per Mbp, little more than half of the rest of the genome. Chromosome 5 also has the highest coverage by LTR retrotransposons (28.3%) and has the youngest intact *Gypsy* elements (1.37 MY vs. 1.54 – 1.64 MY for the other chromosomes). It also has the lowest density of solo LTRs, with a ratio of intact to solo elements of 0.89 compared to 2.6 for whole genome. Thus, unlike the rest of the Brachypodium genome, Bd5S is gaining retrotransposons by replication and losing comparatively fewer by recombination. The syntenic regions of rice (Os4S) and sorghum (Sb6S) also show a low gene density, demonstrating the maintenance of a high repeat content for ~60-70 MY despite the relatively rapid turnover of retroelements (Figure 3A; Figure S15) (*47*). Bd5 is also different from other chromosomes in not having undergone any detectable fusions during its evolution (Figures 4C, E and F). The corresponding chromosome arms in rice (Os4S) and sorghum (Sb6S) (Figures 4C, S15) also have the lowest proportion number of collinear genes; only 72 (~18%) of the 402 genes are conserved in all three species, in contrast to the rest of the genome where approximately 50% of genes are collinear. Bd5S also shows several large rearrangements such as inversions and translocations, in contrast to Bd5L that contains only few rearrangements and many collinear genes (Figure S15). In the Triticeae frequent retroelement insertion and inter-element recombination have deleterious effects on gene islands (*45*). We propose that the ancestral chromosome to Bd5S reached a tipping point where high retrotransposon density had deleterious effects on genes. Bd5S may therefore be a useful microcosm for understanding genome expansion in larger grass genomes.

Small RNA analysis

Endogenous small RNA (~21-24 nt) are non-coding RNA molecules that function in regulation of gene expression, genome defense, and silencing of repeated sequences. Several small RNA classes are important for development, homeostasis, and response to stress (*48*). We analyzed small RNA populations from inflorescence tissues with deep Illumina sequencing (*17*) and mapped them onto the genome sequence (Figure 6, Figure S16, and Table S16). Small RNA reads were most dense in regions of high repeat density, such as centromeres, and lower in regions of high gene density, similar to the distribution reported in Arabidopsis (*49*). Using a modified algorithm to identify phasing patterns of trans-acting (ta-) siRNAs, we identified a total of 413 loci of 19-25 nt small RNAs, of which 198 were 24 nt phased loci. Using the same algorithm just 5 ta-siRNA were identified in Arabidopsis, and none were 24 nt phased. The biological functions of these clusters, which account for a significant number of small RNAs that map outside repeat regions, are currently not known.

Discussion

Research in grasses requires the urgent development of experimental systems for optimizing grass crops for food, feed and fuel production. Brachypodium shows exceptional promise as an experimental organism in the same way that Arabidopsis is an excellent model for dicots. This has led the rapidly expanding Brachypodium research community to develop many genetic and genomic resources that will provide researchers with an unprecedented new opportunity for biological discovery in the grasses. Thus the sequence and analysis of the Brachypodium genome reported here is an important advance towards securing sustainable supplies of food, feed and fuel from new generations of grass crops.

The Brachypodium genome sequence, in combination with those from two other diverse grass subfamilies, enabled reconstruction of chromosome evolution across the broad diversity of grasses. This pan-grass genome sequence 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 centromeric regions. The relatively small genome of Brachypodium contains

many active retroelement families that can contribute to extreme genome size (*24*), but recombination appears to keep genome expansion largely in check. However, the short arm of chromosome 5 deviates from the rest of the genome and exhibits a trend toward genome expansion through increased retroelement numbers and disruption of gene order which are more typical of larger genomes of closely related grasses.

As the first genome sequence of a pooid grass, the Brachypodium genome is aiding the genome-wide interpretation of gene content in the large and complex genomes of wheat and barley, two other pooid grasses that are among the world's most important agricultural species, and in cool season forage and turf crops. For example, the Brachypodium sequence is facilitating map-based gene cloning projects and forming syntenic templates for assembling pooid genome sequences. The overall similarity of Brachypodium, rice and sorghum in terms of gene content and gene family structure indicates the value of Brachypodium as a functional genomics model for all grasses, including those being developed as biomass crops.

3949 words total to here

Acknowledgements

This paper is dedicated to the memory of Mike Gale, who identified the importance of conserved gene order in grass genomes.

This work was supported by the US Department of Energy Joint Genome Institute Community Sequencing Program Contract 776757 to J.D.V, D.F.G., T.C.M. and M.W.B., BBSRC grant BB/E004725/1 to M.W.B. and BBSRC Research Studentship BBS/S/E/2006/13204 to J.W. Work at MIPS was supported by funds from EU Triticeae Genome (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. Manual annotation was supported by USDA NRI Grant 2008-35600-18783 to J.H.C., 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.

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.

Author List and Contribution

International Brachypodium Initiative (Participants are listed under area of contribution, and then by institution).

Principal Investigators: USDA-ARS Western Regional Research Centre John P. Vogel[4]; **USDA-ARS University of Minnesota** David F. Garvin[5]; **Oregon State University** Todd C. Mockler; **John Innes Centre** Michael W. Bevan.

DNA Sequencing and Assembly: Joint Genome Institute Erika Lindquist[9], Igor Grigoriev[9], Hope Tice[9], Mei Wang, Kerrie Barry[9]; **Hudson Alpha Institute** Jeremy Schmutz, Jane Grimwood; **John Innes Centre** Neil McKenzie, Michael W. Bevan

Genome Alignment: USDA-ARS Western Regional Research Centre Naxin Huo, Yong Q. Gu, Mingcheng Luo, Jan Dvorak, Olin D. Anderson, Yong Q. Gu, John P. Vogel; **John Innes Centre** Jonathan Wright, Melanie Febrer, Michael W. Bevan; **University of Silesia** Dominika Idziak, Robert Hasterok.

Transcriptome Sequencing and Analysis: Joint Genome Institute Erika Lindquist[9], Mei Wang; **Oregon State University** Samuel E. Fox[1] Henry D. Priest[1] ,Sergei A. Filichkin1, Scott A.

Givan[1], Douglas W. Bryant[1], Jeff H. Chang[1] , Todd C. Mockler[1;] **Iowa State University** Haiyan Wu[1,2,3] , Wei Wu[1] , An-Ping Hsia[1], Patrick S. Schnable[1] ; **Washington State University** Anantharaman Kalyanaraman[4], **University of Florida** Brad Barbazuk[5], **Waksman Institute** Todd P. Michael[2] , **University of Massachusetts** Samuel Hazen[3] , **Western Regional Research Center**, **USDA** John Vogel[4], Jennifer Bragg[4], Debbie Laudencia[4], **USDA-ARS Plant Science Research Unit and University of Minnesota** David F. Garvin[5], **University of Wisconsin** Yiqun Weng[6] **John Innes Centre** Neil McKenzie, Michael W. Bevan.

Gene Analysis and Annotation: Helmholtz Zentrum München Georg Haberer[2], Manuel Spannagl[2], Klaus Mayer[2]; **Technical University München** Thomas Rattei. **UC Berkeley** Therese Mitros[3], Dan Rokhsar[3];

Repeats Analysis: University of Zurich Thomas Wicker, Jan P. Buchmann; **MTT Agrifood Research Finland and University of Helsinki** Jaakko Tanskanen, Alan H. Schulman; **Helmholtz Zentrum München** Heidrun Gundlach; **John Innes Centre** Jonathan Wright; **Federal University of Pelotas** Antonio Costa de Oliveira, Luciano Carlos da Maia. **USDA-ARS Western Regional Research Centre** Yong Gu; **Michigan State University** Ning Jiang; **China Agricultural University** Jinxin Lai; **Purdue University** Jianxin Ma.

Comparative Genomics: UMR INRA-UBP 1095 Clermont-Ferrand Jerome Salse, Florent Murat, Michael Abrouk;; **Helmholtz Zentrum München** Georg Haberer[2], Manuel Spannagl[2], Klaus Mayer[2] ;**Rutgers University** Remy Bruggmann; **UC Davis** Jan Dvorak.

Small RNA Analysis: Oregon State University Noah Fahlgren[1], Samuel E. Fox[1] Christopher M. Sullivan[1], Todd C. Mockler, James C. Carrington[1]; **UC San Diego** Elisabeth Chapman[2]; **University of Delaware** Jixian Zhai[3],Matthias Ganssmann[3], Guna Sai Ranjan[3], Marcelo German[3], Greg D. May[4], Blake C. Meyers[3], Pamela J. Green[3].

Manual Annotation and Gene Family Analysis: **USDA-ARS Western Regional Research Centre** John P. Vogel; **VIB Gent** Stephanie DeBodt, Wim Verelst, Dirk Inze, **Institut de Biologie Moleculaire des Plants du CNRS** Maren Heese, Arp Schnittger; **University of Massachusetts** Sam Hazen; **Université de Picardie Jules Verne** Jerome Pelloux; **Illinois State University** John Sedbrook, M. Cass; **University of California, Davis** Pam Ronald;

Oregon State University Jim Carrington Noah Fahlgren, Samuel Fox, Sergei Filchin, Henry D. Priest, Todd C. Mockler, Jeff H. Chang, Jeffrey A. Kimbel; **USDA-ARS Plant Science Research Unit and University of Minnesota** David F. Garvin[5], **China Agricultural University**: Yu Cui, Shuhong Ouyang, Qixin Sun, Zhiyong Liu. Not complete

Affiliations of Participants: DOE Joint Genome Institute, Walnut Creek, CA 94598, USA; John Innes Centre, Norwich NR4 7UJ, UK; University of Silesia, 40-032 Katowice, Poland; HudsonAlpha Institute, Huntsville, AL 35806, USA; Oregon State University, Corvallis, OR 97331, USA; Iowa State University, Ames, IA 50011, USA; Washington State University, Pullman, WA 99163, USA; University of Florida, Gainsville, FL 32611, USA; Waksman Institute, Rutgers University, Piscataway, NJ 08855-0759, USA; University of Massachusetts, Amherst, MA 01003-9292, USA; Western Regional Research Centre, USDA, Albany, CA 94710, USA; USDA-ARS Plant Science Research Unit and University of Minnesota , St. Paul, MN 55108-6026, USA; University of Wisconsin, Madison, WI 53726, USA; Helmholtz Zentrum München, D-85764 Neuherberg, Germany; Technical University München, 80333 München, Germany; University of California, Berkeley, CA 94720, USA; University of Zurich, 8008 Zurich, Switzerland; MTT Agrifood Research and University of Helsinki, FIN-00014 Helsinki, Finland; Federal University of Pelotas Pelotas, 96001-970,RS, Brazil; Michigan State University, East Lansing, MI 48824, USA; China Agricultural University, Beijing 100094, China; Purdue University, West Lafayette, IN 47907, USA; Institut National de la Recherché Agronomique UMR 1095, 63100 Clermont-Ferrand, France; University of California Davis, Davis, CA 95616, USA; University of California San Diego, La Jolla, CA 92093, USA; University of Delaware, Newark, DE 19716, USA;

References

1. P. R. Crane, E. M. Friis, K. R. Pedersen, *Nature* **374**, 27 (1995).
2. F. Darwin, A. C. e. Seward, *More Letters of Charles Darwin* (John Murray, London, 1903), pp.
3. E. A. Kellogg, *Plant Physiol* **125**, 1198 (2001).
4. C. Somerville, *Science* **312**, 1277 (2006).
5. International, Rice, Genome, Sequencing, Project, *Nature* **436**, 793 (2005).
6. A. H. Paterson *et al.*, *Nature* **457**, 551 (2009).
7. F. Wei *et al.*, *PLoS Genet* **3**, e123 (2007).
8. G. Moore, K. M. Devos, Z. Wang, M. D. Gale, *Curr Biol* **5**, 737 (1995).
9. J. Draper *et al.*, *Plant Physiol* **127**, 1539 (2001).
10. J. Vogel, T. Hill, *Plant Cell Rep* **27**, 471 (2008).
11. P. Vain *et al.*, *Plant Biotechnol J* **6**, 236 (2008).
12. D. F. Garvin *et al.*, *Crop Science* **48**, S69 (2008).
13. D. B. Jaffe *et al.*, *Genome Res* **13**, 91 (2003).
14. M. D. Bennett, I. J. Leitch, *Ann Bot (Lond)* **95**, 45 (2005).
15. J. P. Vogel, D. F. Garvin, O. M. Leong, D. M. Hayden, *Plant Cell, Tissue and Organ Culture* **84**, 199 (2006).
16. R. Hasterok, J. Draper, G. Jenkins, *Chromosome Res* **12**, 397 (2004).
17. supporting, online, text.
18. T. Tanaka *et al.*, *Nucleic Acids Res* **36**, D1028 (2008).
19. S. Fox, S. Filichkin, T. Mockler, Eds., *Applications of ultra high throughput sequencing*, vol. 553 (Humana Press, 2009).
20. www.brachybase.org; http://mips.helmholtz-muenchen.de/plant/brachypodium/index.jsp;www.modelcrop.org
21. J. Gray *et al.*, *Plant Physiol* **149**, 4 (2009).
22. G. B. Fincher, *Curr Opin Plant Biol* **12**, 140 (2009).
23. N. C. Carpita, *Annu Rev Plant Physiol Plant Mol Biol* **47**, 445 (1996).
24. J. L. Bennetzen, E. A. Kellogg, *Plant Cell* **9**, 1509 (1997).
25. C. Vitte, O. Panaud, H. Quesneville, *BMC Genomics* **8**, 218 (2007).
26. T. Wicker, B. Keller, *Genome Res* **17**, 1072 (2007).
27. J. S. Ammiraju *et al.*, *Plant J* **52**, 342 (2007).
28. T. Wicker *et al.*, *Plant Physiol* **149**, 258 (2009).
29. T. E. Bureau, S. R. Wessler, *Plant Cell* **6**, 907 (1994).
30. T. E. Bureau, S. R. Wessler, *Proc Natl Acad Sci U S A* **91**, 1411 (1994).
31. N. Jiang, C. Feschotte, X. Zhang, S. R. Wessler, *Curr Opin Plant Biol* **7**, 115 (2004).
32. N. Jiang, Z. Bao, X. Zhang, S. R. Eddy, S. R. Wessler, *Nature* **431**, 569 (2004).
33. M. Morgante *et al.*, *Nat Genet* **37**, 997 (2005).
34. M. Freeling, S. Subramaniam, *Curr Opin Plant Biol* **12**, 126 (2009).
35. Grass Phylogeny Working Group *Annals of the Missouri Botanical Garden* **88**, 373 (2001).
36. B. S. Gaut, *New Phytologist* **154**, 15 (2002).
37. E. Bossolini, T. Wicker, P. A. Knobel, B. Keller, *Plant J* **49**, 704 (2007).
38. M. Charles *et al.*, *Mol Biol Evol* **26**, 1651 (2009).
39. N. Stein *et al.*, *Theor Appl Genet* **114**, 823 (2007).
40. M. C. Luo *et al.*, *Proc. Natnl. Acad. Sci. U.S.A. in press* (2009).
41. L. L. Qi *et al.*, *Genetics* **168**, 701 (2004).
42. J. Salse *et al.*, *Plant Cell* **20**, 11 (2008).
43. A. H. Paterson *et al*., *Proc Natl Acad Sci U S A* **101**, 9903 (2004).
44. M. A. Lysak *et al.*, *Proc Natl Acad Sci U S A* **103**, 5224 (2006).
45. C. M. Vicient, R. Kalendar, A. H. Schulman, *J Mol Evol* **61**, 275 (2005).
46. B. C. Meyers *et al.*,*Plant Cell* **15**, 809 (2003).
47. J. Ma, J. L. Bennetzen, *Proc Natl Acad Sci U S A* **101**, 12404 (2004).
48. O. Voinnet, *Cell* **136**, 669 (2009).
49. R. Rajagopalan, H. Vaucheret, J. Trejo, D. P. Bartel, *Genes Dev* **20**, 3407 (2006).
50. T. Wicker *et al.*, *Nat Rev Genet* **8**, 973 (2007).

Figure Legends

Figure 1. Chromosomal distribution of the main Brachypodium genome features.

The abundance and distribution of the following main genome elements are shown. Complete LTR retroelements (cLTR); solo-LTRs (sLTR); autonomous DNA transposons (DNA-TEs); deletion derivatives of DNA transposons (MITES); gene exons (CDS); gene introns and satellite tandem arrays (CEN) are shown. The bar-charts are from 0 to 100 percent bp coverage of the respective window. The heat map tracks have different scales: CEN [0-55|scaled to max10] %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.

Figure 2. Gene identification and distribution among three grass subfamilies

A. Coverage over the length of Bradi1.0 gene features. 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. Box-and-whisker plots of Illumina coverage calculated as the percentage of bases along the length of the sequence feature that was supported by Illumina reads for 5' untranslated regions (5UTR), 3' untranslated regions (3UTR), introns, exons, genes, cDNAs, coding sequences (CDS), and splice junctions (SJs). The bottom and top of the box represent the 25th and 75th quartiles, respectively. The white line is the median and the open red diamonds are the mean.

B. The secreted proteomes of Arabidopsis, rice and Brachypodium were identified by predicting N- terminal signal peptides (SP) 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.

C. Venn diagram showing the distribution of shared gene families between three major subfamilies of grasses. The *Erhartoideae* (complete rice RAP2 gene predictions), *Panicoideae* (Sorghum V1.4 gene predictions) and *Pooideae* (Brachypodium v1.0 gene predictions, and

Triticum aestivum and *Hordeum vulgare* TCs/EST sequences). Paralogous gene families were collapsed in these datasets. The three grass subfamilies shared 77%-84% of the gene families. The *Pooideae*-specific gene family set contains only 265 gene families comprising 1636 genes.

Figure 3. Retro- and DNA- transposons in the Brachypodium genome.

A. Retroelement family ages. The age distribution and frequency of intact *Copia* and *Gypsy* LTR retrotransposons (green bars) and *Copia* and *Gypsy* solo LTRs (dotted line) grouped in age classes of 0.1 MY. Fitted exponential decay curves for the half-life of intact elements are shown.

B. 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 represents the fused transposase gene.

Figure 4. Brachypodium genome evolution

A. The distribution maxima of mean synonymous substitution rates (Ks) of Brachypodium, rice, sorghum and wheat orthologous gene pairs (figure S13) were used to define the divergence times of these species and the age of inter- chromosomal duplications in Brachypodium. WGD=

Whole Genome Duplication. The numbers refer to the predicted divergence times measured as millions of years ago (MYA).

B. Diagram showing the six major inter-chromosomal duplications, defined by 723 paralogous relationships, as coloured bands linking the five chromosomes.

C. Identification of precise chromosome relationships between the Brachypodium, rice, and Sorghum genomes. Orthologous relationships between the 25,532 protein-coding Brachypodium genes, 7,216 sorghum orthologs (12 syntenic blocks), 8,533 rice orthologs (12 syntenic blocks) were defined. Sets of collinear orthologous relationships are represented by a coloured band according to each Brachypodium chromosome. Each Brachypodium chromosome is represented by a distinguishing colour (blue- chr. 1; yellow- chr.2; violet- chr.3; red-chr.4; green- chr.5). The white region in each chromosome represents the centromeric region. A twist represents an inversion of order.

D. The patterns of collinear orthologous gene relationships with Brachypodium chromosomes can be interpreted as nested insertions. The diagram, which is not drawn to scale, shows how 12 rice chromosomes can form 5 Brachypodium chromosomes. The dot represents the centromeric region.

E. Orthologous gene relationships between Brachypodium and the pooid grasses barley and *Ae.tauschii* were aligned according to genetically-mapped ESTs (barley 1,015 ESTs, *Ae. tauschii* 863 ESTs). 2,516 orthologous relationship defined 12 syntenic blocks. These are shown as coloured bands.

F. Orthologous gene relationships between Brachypodium and hexaploid bread wheat defined by 5,003 ESTs mapped to wheat deletions. Each set of orthologous relationships is represented by a band that is evenly spread across each deletion interval on the representations of wheat chromosomes. As the relative order of genes within each wheat deletion interval are not yet known, thus the connecting bands cannot be oriented.

Figure 5. A recurring pattern of nested chromosome fusions in grasses.

A. The colors define each rice chromosome (Os1-Os12) on which the closest Brachypodium homolog is located. Large syntenic regions are revealed by the predominant color. Chromosomes descended from an ancestral chromosome (A4-A11) through whole genome duplication are displayed in shades of the same color. Gene density is indicated as a red line above the chromosome maps and was calculated in sliding windows of 1 Mbp with a step of 100,000 bp. Major discontinuities in gene density mark syntenic breakpoints, which are marked by a diamond.

B. A pattern of nested insertions for Brachypodium chromosomes 1 through 4 can explain the observed syntenic break points. All nested insertions targeted the centromeric region, even in very asymmetric chromosomes (e.g. Bd4). Bd5 has not undergone chromosome fusion.

C. Examples of nested chromosome fusions in *Sorghum bicolor* (Sb) chromosomes 1 and 2.

D. Schematic representation of identifiable nested chromosome fusions in barley inferred from genetic mapping data. Nested insertions were not identified in other chromosomes, possibly due to the low resolution of genetic markers.

Figure 6. Genome-wide distribution of small RNA, genes and repeat elements in the *Brachypodium distachyon* genome.

Each Brachypodium chromosome (1-5) is shown as an ideogram at the top of the Figure. Total small RNA reads (black lines) and total small RNA loci (red lines) are shown on the top panel. Histograms plot 21nt (blue) or 24 nt (red) small RNA reads normalized for repeated matches to the genome, respectively. The 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 (green), normalized for ambiguous matches to the genome. The 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 = 100000 nt, scroll = 20000 nt).

Feature	Rice (RAP2)	Brachypodium (v1.0)	Sorghum (v1.4)	Athal (TAIR8)
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 Scaffolds (bp)	---	774,881	79,311,565	---
Loci (protein coding)	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

Table 1 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).

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

[2] Only *bona fide* gene models of sorghum were considered for this table (*6*).

	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

Table 2. Brachypodium transposable element content. The table summarizes the annotation of full length elements and transposon fragments that were classified according to (*50*) .

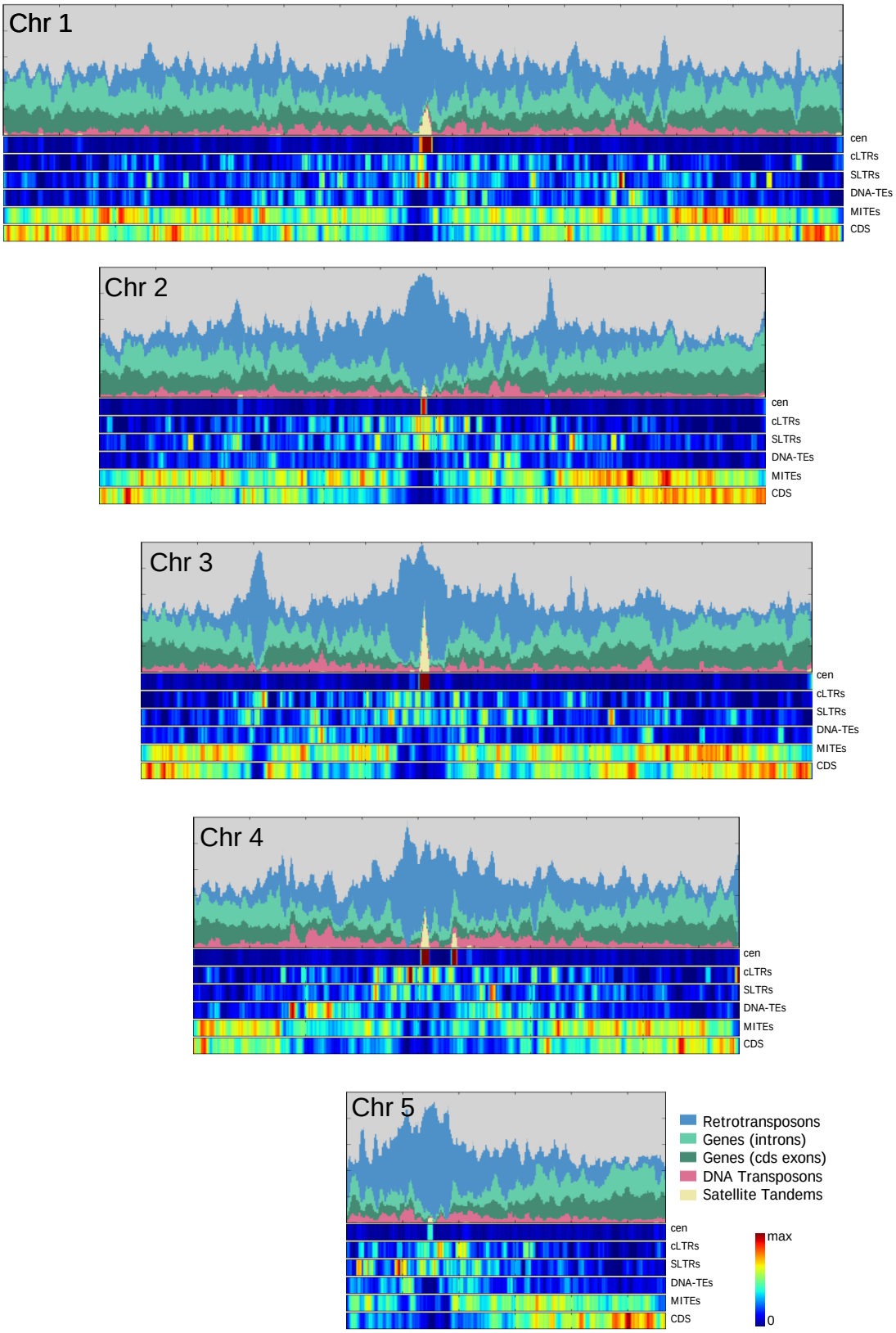
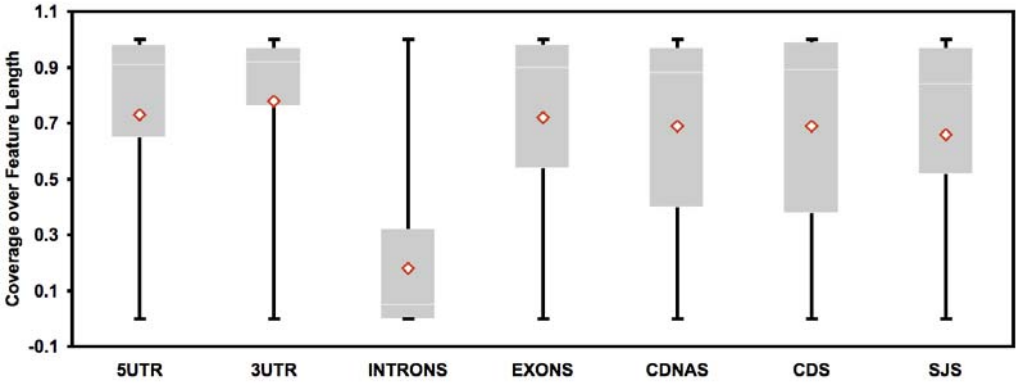
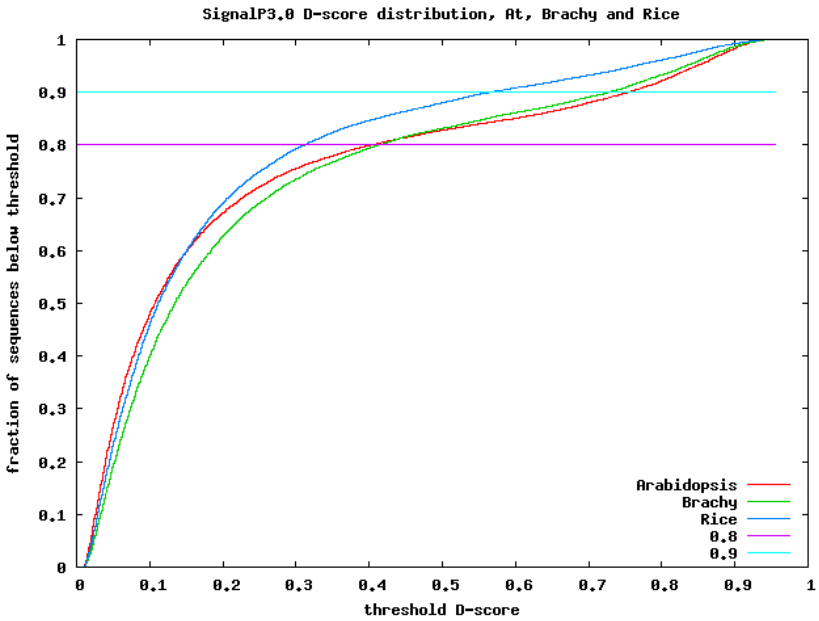

Chr 1
cen
cLTRs
SLTRs
DNA-TEs
MITEs
CDS
Chr 2
cen
cLTRs
SLTRs
DNA-TEs
MITEs
CDS
Chr 3
cen
cLTRs
SLTRs
DNA-TEs
MITEs
CDS
Chr 4
cen
cLTRs
SLTRs
DNA-TEs
MITEs
CDS
Chr 5
cen
cLTRs
SLTRs
DNA-TEs
MITEs
CDS
Retrotransposons
Genes (introns)
Genes (cds exons)
DNA Transposons
Satellite Tandems
max
0

Figure 1, IBI

A



B



C

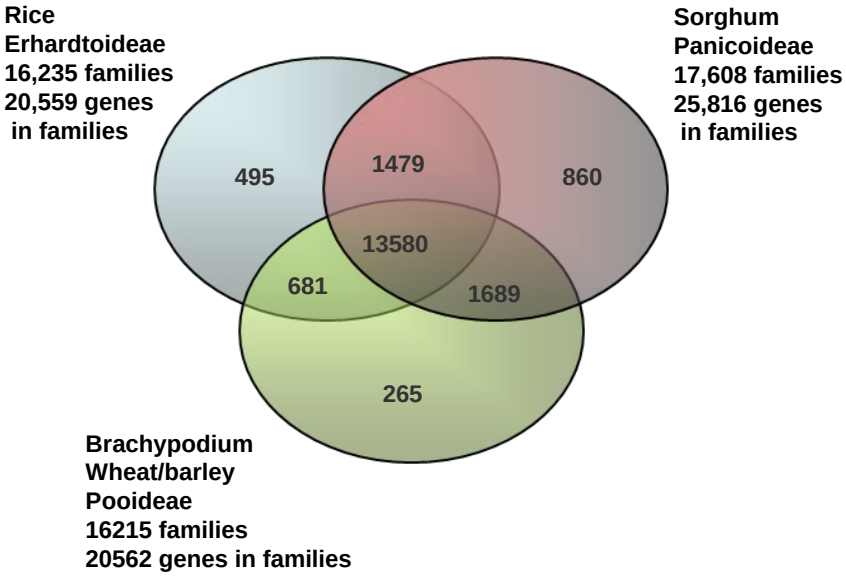
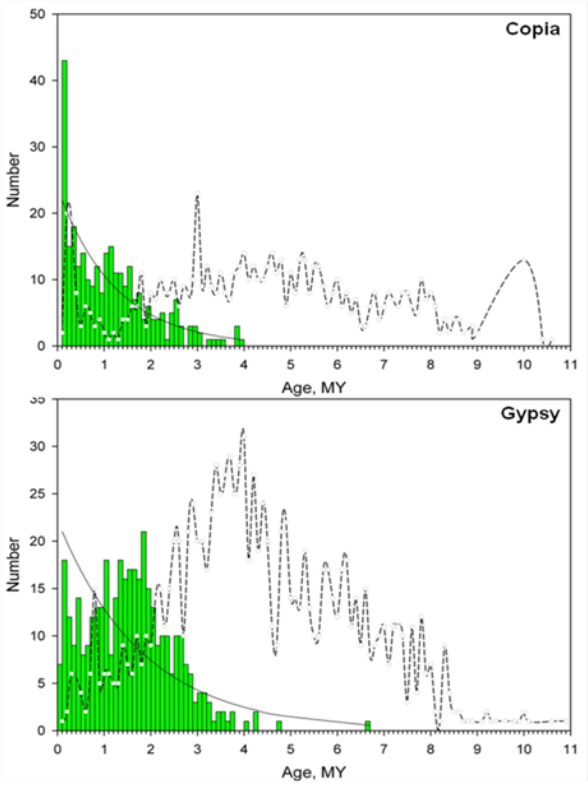


Figure 2, IBI

A



B

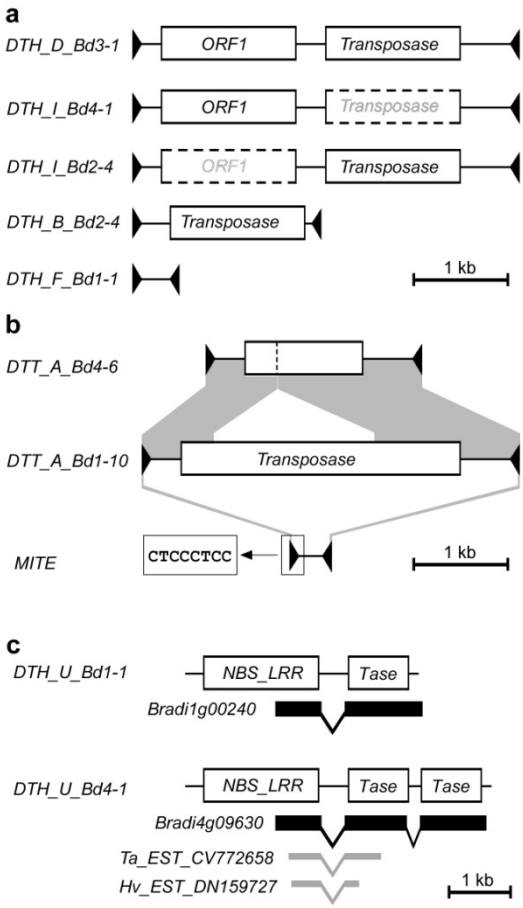


Figure 3, IBI

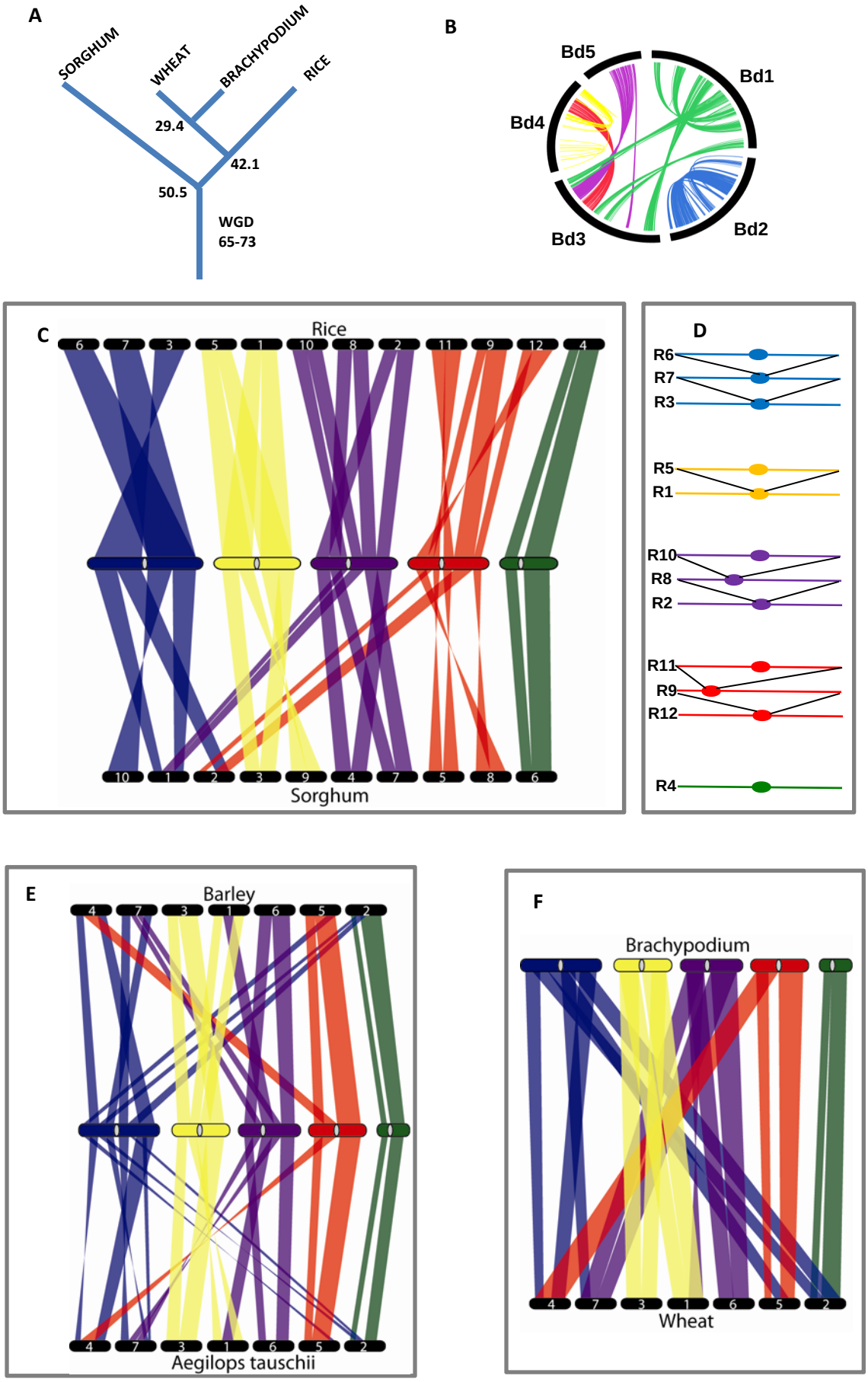

A
SORGHUM
WHEAT
BRACHYPODIUM
RICE
29.4
42.1
50.5
WGD
65-73
B
Bd5
Bd1
Bd4
Bd3
Bd2
C
Rice
6 7 3 5 1 10 8 2 11 9 12 4
10 1 2 3 9 4 7 5 8 6
Sorghum
D
R6
R7
R3
R5
R1
R10
R8
R2
R11
R9
R12
R4
E
Barley
4 7 3 1 6 5 2
4 7 3 1 6 5 2
Aegilops tauschii
F
Brachypodium
4 7 3 1 6 5 2
Wheat

Figure 4 IBI

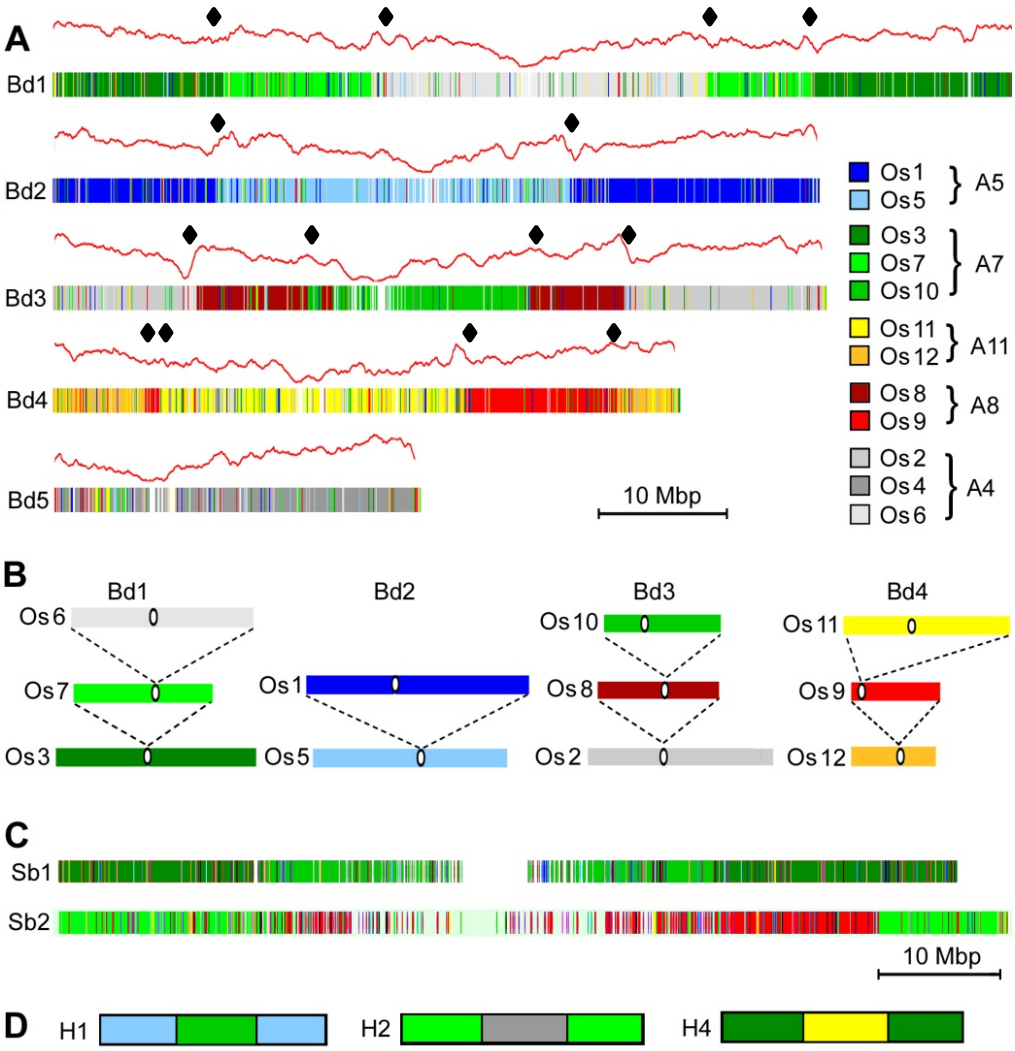

A
Bd1
Bd2
Bd3
Bd4
Bd5
10 Mbp
Os1
Os5
A5
Os3
Os7
Os10
A7
Os11
Os12
A11
Os8
Os9
A8
Os2
Os4
Os6
A4
B
Bd1
Bd2
Bd3
Bd4
Os6
Os7
Os3
Os1
Os5
Os10
Os8
Os2
Os11
Os9
Os12
C
Sb1
Sb2
10 Mbp
D
H1
H2
H4

Figure 5, IBI

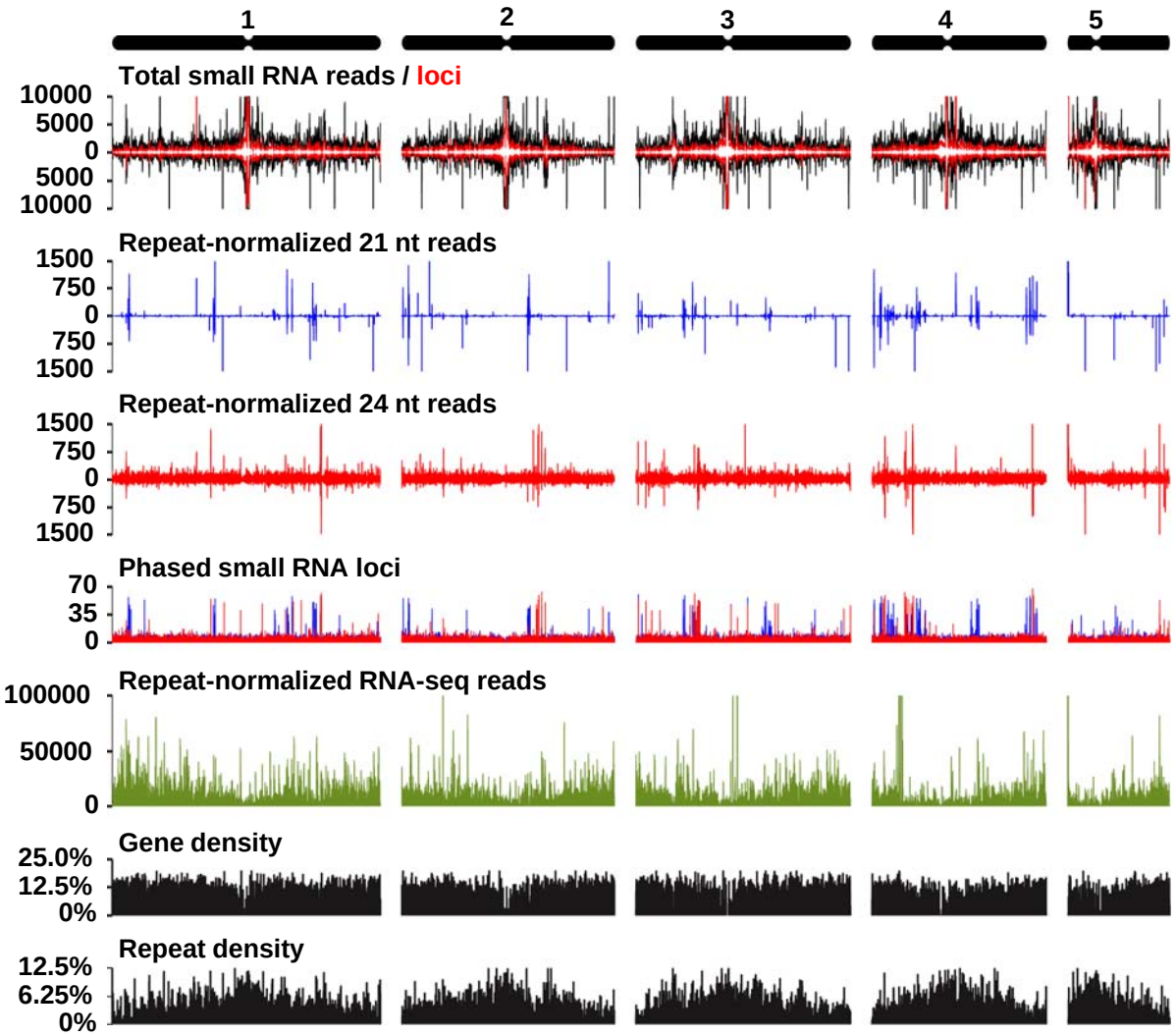

1
2
3
4
5
Total small RNA reads / loci
10000
5000
0
5000
10000
Repeat-normalized 21 nt reads
1500
750
0
750
1500
Repeat-normalized 24 nt reads
1500
750
0
750
1500
Phased small RNA loci
70
35
0
Repeat-normalized RNA-seq reads
100000
50000
0
Gene density
25.0%
12.5%
0%
Repeat density
12.5%
6.25%
0%

Figure 6, IBI

Supplementary Information

S1. 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*) (*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 (Table S1) and sequenced using standard Sanger protocols on ABI 3730 xl instruments. The total number of reads from each library is shown in Table S1.

Table S1. Assembly Input

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 [1,2,3] (HinDIII)	103,216	30,704	0.05
BAC DB [1,2,3] (BamH1)	108,177	36,388	0.04
BAC BD_CBa [4] (EcoR1)	124,935	25,948	0.05
BAC BD_ABa [4] (HinDIII)	149,112	34,177	0.07
		4,761,917	9.43

1. (*2*)
2. (*3*)
3. (*4*)
4. (*5*)

Table S2: Raw Assembly Output

Scaffold Length (bp)	Number of Scaffolds	Number of Contigs	Total Scaffold Length (bp)	Total Contig Length (bp)	Coverage
all	217	2,067	272,077,374	272,287,606	99.60%
1,000	208	2,058	272,071,085	272,281,317	99.60%
2,500	193	2,043	272,048,669	272,258,901	99.60%
5,000	127	1,925	271,781,248	272,020,434	99.61%
10,000	60	1,787	271,288,614	271,563,788	99.62%
25,000	20	1,711	270,712,788	271,003,970	99.63%
50,000	13	1,684	270,471,535	270,814,201	99.65%
100,000	11	1,671	270,362,712	270,737,212	99.66%
250,000	11	1,671	270,362,712	270,737,212	99.66%
500,000	11	1,671	270,363,712	270,737,212	99.66%
1,000,000	10	1,665	269,833,561	270,190,573	99.66%
2,500,000	10	1,665	269,833,561	270,190,573	99.66%
5,000,000	10	1,665	269,833,561	270,190,573	99.66%

Table S3. Final Genome Release

Final Contigs	1,630 contigs
Total Genome Size	271,148,425 bp
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	270.8 Mb
Release Scaffold N/L50	3/59.3 Mb
Release Contig N/L50	252/347.8 Kb
Final Genome Coverage	9.4x

Organelle DNA in the nuclear genome

A total of 1,131 chloroplast DNA 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. Essentially all inserts were less than 0.5 kb, but 17 chloroplast insertions contained intact genes, and approximately 20% of chloroplast and 8% of mitochondrial inserts were identical to organelle sequences, indicating ongoing insertion events.

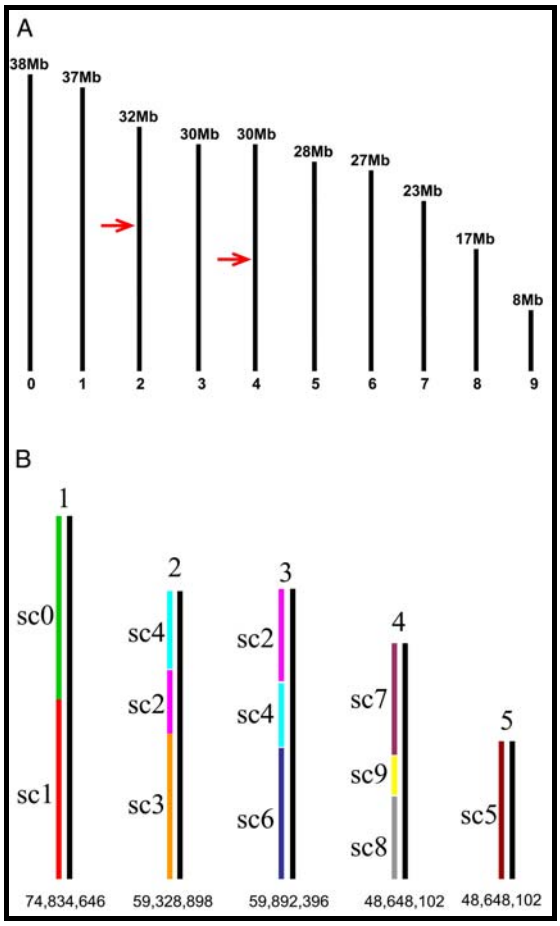


Figure S1. Ordering scaffolds using a genetic map. To verify and assemble the 8x scaffolds into chromosome-scale assemblies we compared the scaffolds to a high-density genetic map constructed from 562 SNP markers selected to be evenly spaced along the 4X scaffolds (full details of the map will be published elsewhere). (A) Only two false joins were detected and they were broken where indicated by red arrows. (B) Color coded assignment of scaffolds to the five Brachypodium chromosomes.

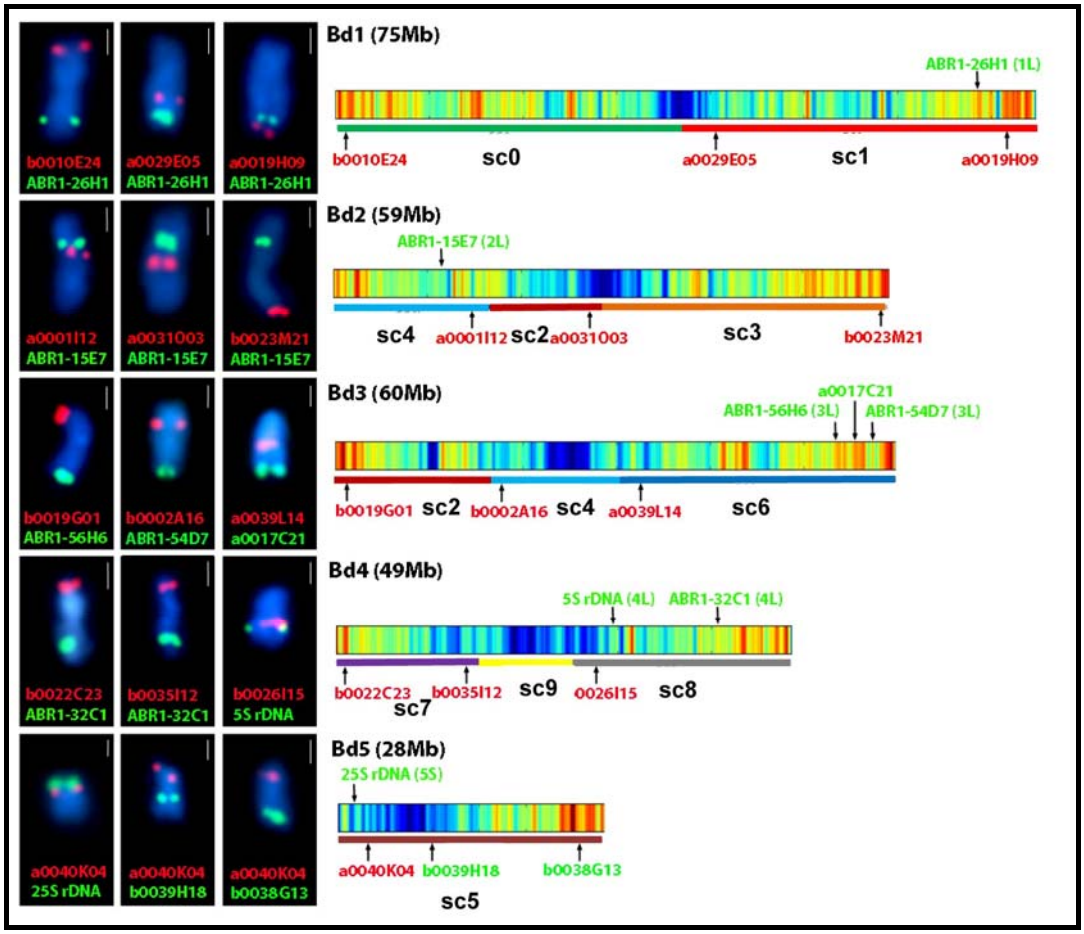


Figure S2. Aligning genome sequence assemblies to Brachypodium chromosomes. Supercontigs from the sequence assemblies were aligned to the Brachypodium karyotype using fluorescently labelled BACs from a physical map integrated with the sequence assemblies (*5*). The methods used are described in S9 below. 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 supercontigs (SC) 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 supercontigs are colored according to Figure S1.

Table S4. EST Resources used for genome annotation

LibraryName	#ESTs	Platform	Sequenced by	Bd genotype	Tissue/Stage/Treatment etc...	Normalization	Contributor/Reference
CCXU	49540	454	JGI	Bd21	callus	N/A	Vogel, Bragg
CFAA	948	454	JGI	Bd21	roots	DSN	Garvin
CFAB	234	454	JGI	Bd21	developing seeds	DSN	Mockler, Michael, Laudencia-Chingcuanco
CFAC	1851	454	JGI	Bd21	diurnally sampled whole seedlings	DSN	Mockler
CFCF	405974	454	JGI	Bd21	diurnally sampled roots	DSN	Garvin
CFCG	317095	454	JGI	Bd21	diurnally sampled leaves + stems	DSN	Mockler
CFCH	362432	454	JGI	Bd21	diurnally sampled flowers RNA	DSN	Mockler
CFCI	253491	454	JGI	Bd21	callus	DSN	Vogel, Bragg
CFFH	129769	454	JGI	Bd21	diurnally sampled leaves + stems + callus	DSN	Mockler, Vogel, Bragg
CFFI	139968	454	JGI	Bd21	diurnally sampled leaves + stems + callus	DSN	Mockler, Vogel, Bragg
CFFN	93222	454	JGI	Bd21	diurnally sampled leaves + stems + callus	DSN	Mockler, Vogel, Bragg
AC60	170521	454	Schnable	PI 185133 (source of Bd2-3)	root tips	N/A	Schnable
AC61	89277	454	Schnable	PI 185134 (source of Bd3-1 and 3-2)	root tips	N/A	Schnable
AC63	157349	454	Schnable	PI 245730 (source of Bd18-1)	root tips	N/A	Schnable
AC64	122320	454	Schnable	PI 254867 (source of Bd21)	root tips	N/A	Schnable
CCXF	25494	Sanger	JGI	Bd21	abiotic stress + biotic stress	DSN	Mockler, Chang, Hazen, Weng
CCXG	28229	Sanger	JGI	Bd21	superpool	DSN	Mockler, Vogel, Hazen, Chang, Michael, Garvin, Bevan
CCYO	26237	Sanger	JGI	Bd21	flower + flower drought	DSN	Bevan
CCYP	27821	Sanger	JGI	Bd21	leaf+ leaf drought	DSN	Bevan
callus	4196	Sanger	Vogel	Bd21	callus	N/A	Vogel
leaf	3780	Sanger	Vogel	Bd21	leaf	N/A	Vogel
root	3869	Sanger	Vogel	Bd21	root	N/A	Vogel
seed	4688	Sanger	Vogel	Bd21	seed	N/A	Vogel
stem	3907	Sanger	Vogel	Bd21	stem	N/A	Vogel
SuperPool	289000000	Illumina	Mockler	Bd21	superpool	DSN	Mockler, Vogel, Hazen, Chang, Michael, Garvin, Bevan, Laudencia-Chingcuanco, Weng
Total 454:	2293991						
Total Sanger:	128221						
Total Illumina	~289M						

S2. 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 (*7*) 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) (*8, 9*), sorghum (version 1.4) (*10*) and Arabidopsis (version TAIR8) (*11, 12*) 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. Table S4 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 (*13*) 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 initial Arabidopsis convention (*14*) 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 complied from SWISSPROT, rice (RAP2 and TIGR5), sorghum (version 1.4) and Arabidopsis (TAIR8) protein databases. 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 low experimental support (classes 0 and 1, Figure S2) were independently evaluated for transcriptional evidence using 10.2 Gb Illumina transcriptome data (Section S3) and only genes with at least 20% coverage over the length of the predicted cDNA by Illumina data were retained, as described below.
tRNA genes were identified by tRNA-SEscan (*15*) 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.

S3. Illumina Transcriptome Methods

Full-length enriched (FL) and randomly primed (RP) cDNA libraries were prepared from RNA isolated as described in Table S4, and sequenced using an Illumina 1G Genome Analyzer essentially as described (*16*). 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 and only reads with a length of exactly 32 bases were retained for subsequent analysis. 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.

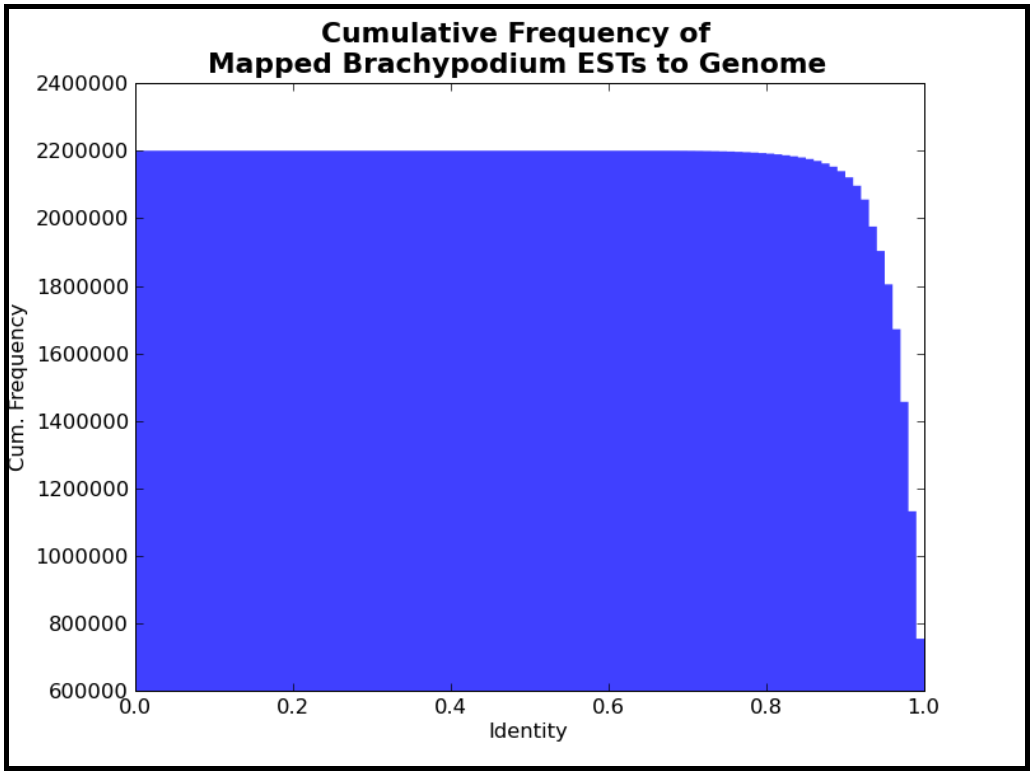


Figure S3. Mapping of Brachypodium ESTs onto the Genomic Sequence.

Brachypodium ESTs/TCs were anchored onto the genomic assemblies as optimal spliced alignments using the program GenomeThreader. In total, 2,200,497 out of 2,305,135 transcript sequences (95.5%) could be mapped to the genomic sequence with a minimum alignment length of 50% of the transcript size. On the y-axis, the cumulative frequency of anchored ESTs/TCs is shown according to its dependence of alignment identity on the x-axis. For each EST/TC, the highest alignment identity has been selected in case of several genomic alignment positions. The large majority of ESTs/TCs could be mapped with high sequence identities, ≥1,900,000 and ≥2,100,000 sequences with an identity ≥95% and ≥90%, respectively.

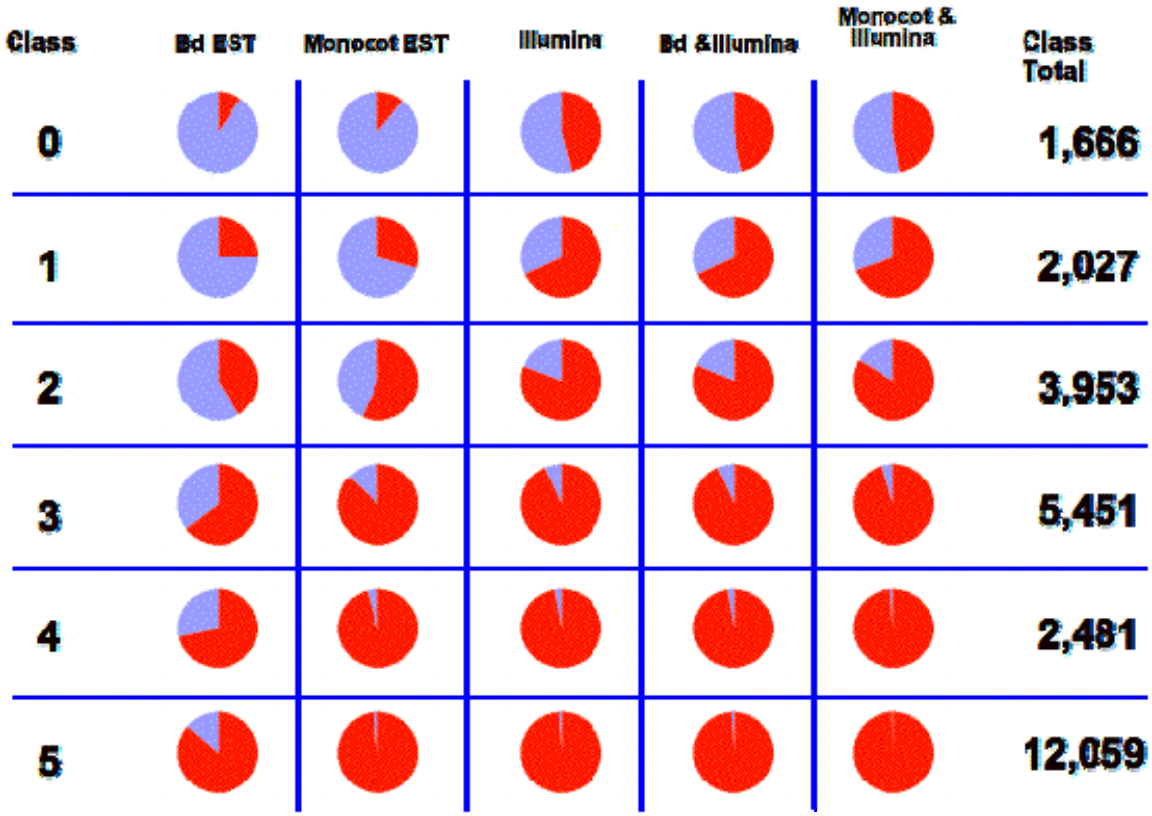


Figure S4. Class distribution and extrinsic evidence for Brachypodium gene predictions.

Initial Brachypodium gene predictions (v1.0) 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 dataset

S3 Manual annotation and gene family analysis

Gene models (2,755) from gene families or pathways of potential relevance to bioenergy research were selected for manual annotation based on BLAST scores to known genes and/or from the presence of pfam domains (Table S5). Selected genes were manually examined using EST alignments, Illumina transcriptome data, splice site verification by Illumina sequence and alignment to previously described genes from other organisms, and edited. Phylogenetic analysis of 62 gene families demonstrated that most cases Brachypodium, rice and Sorghum had very similar gene family compositions, with the exception of flowering time, small RNA processing, Receptor-like Ser–Thr kinases, and cellulose synthase-like genes.

The flowering time pathway is highly conserved and contained the expected Brachypodium genes(*17*) that are shared between Arabidopsis and rice. However, rice utilizes 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 signalling pathways to control *Ehd1* expression (*18*). 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 in this plant; 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 (Figure S5). Therefore this family member may have been lost in *Brachypodium.* However, Brachypodium does have five other RDR genes in the other three RDR clades.

Receptor-like Ser–Thr kinases (RLKs) constitute a major gene family in land plants with respectively over 600 and 800 members encoded in the Arabidopsis and Brachypodium genomes. The CrRLK1L subfamily of plant-specific proteins is defined by the conserved extracellular domain and the presence of an “RD” kinase domain, in contrast to the RLKs involved in non-self recognition (*19*). This family has 17 members in Arabidopsis, 14 in Brachypodium and 20 in rice (Figure S5). 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 (*20*). AmRLK is expressed in the petal epidermis of Antirrhinum and may be involved in the polar outgrowth of the epidermal cells (*21*). The FER subclass which contains a single gene in Arabidopsis has seven members in rice and three in Brachypodium. This might reflect a diversification of pollen tube recognition which 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 absence of petals in grasses.

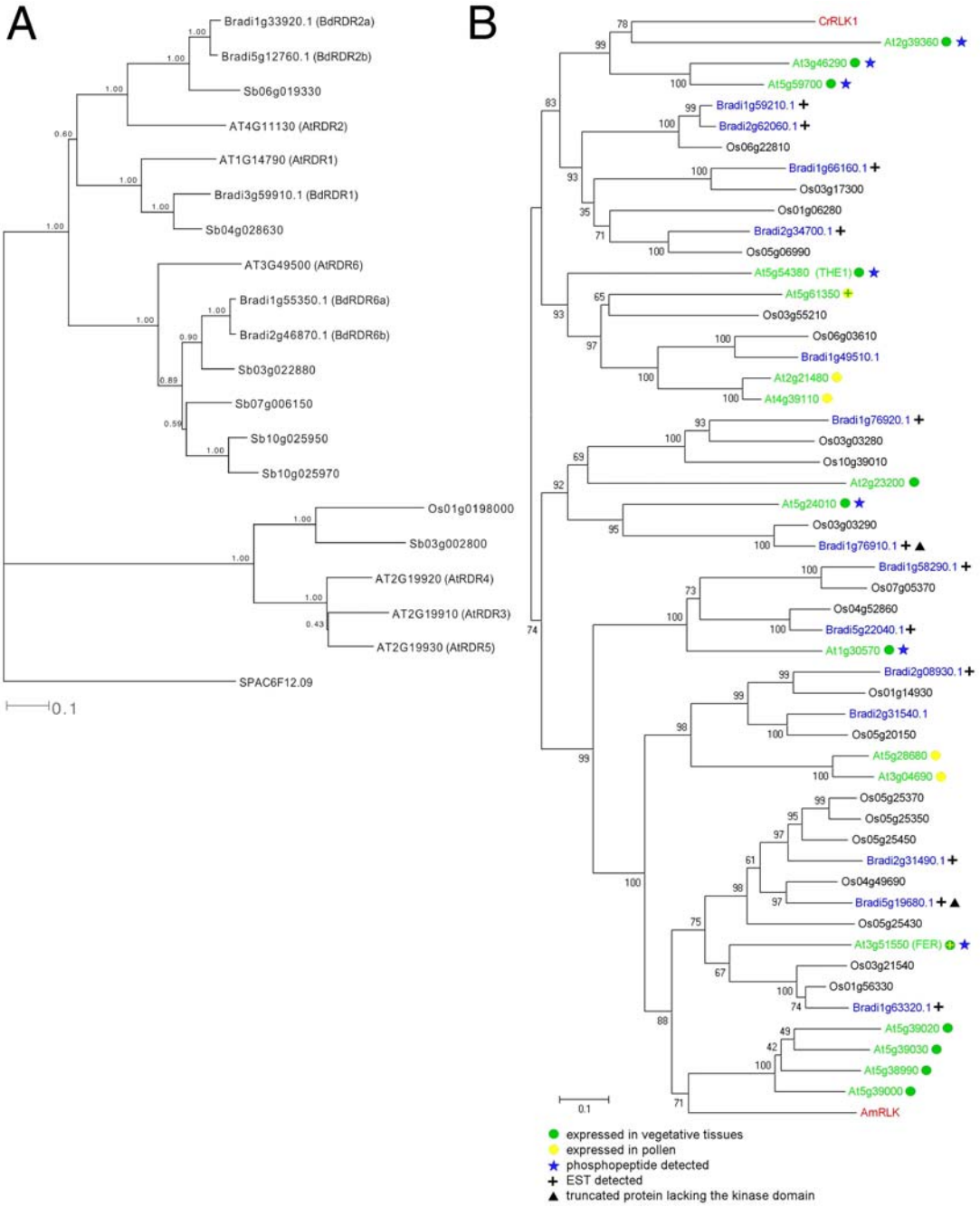


Figure S5. Phylogenetic trees of (A) RDR genes, and (B) CrRLK1L, two gene families with different family composition among the grasses.

Table S5. Manually annotated genes.

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 transcrition factor	transcription factor	109	28
NAC transcrition 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/LFAFY-like	circadian clock/flowering time	1	0
Zea Maize thick tassel dwarf1 (TD1) ortholog[2]	leucine-rich repeat receptor-like kinase	1	0
Zea Maize ramosa2 (ra2) ortholog[2]	transcription factor	1	0
Zea Maize teosintebranched1 (tb1) ortholog[2]	transcription factor	1	0
Zea Maize 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.

CSL TREE GOES HERE I NEED TO FINISH EDITING THE TREE

Figure S6. Consensus neighbor-joining tree of the CSL gene family based on 1,000 bootstrap trees. Note that the grasses have a similar distribution of family members with the exception of CSLJ. After noting that poplar sequences were included in the CSLJ clade, we searched for additional dicot CSLJ genes and added the ### genes from ??????? to the tree. Also note that poplar has two CSL genes that fall between the established CSL? and The Sorghum and poplar gene models were not edited, so there may be additional CSL genes not represented because they were truncated or otherwise mis-annotated.

Table S6. Genes manually assigned to families.

Gene family	Number of genes	general function
Kinase (140 subfamilies)[1]	1,440	phosphorylation
RING	545	protein degradation
F-Box	427[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
sucrose synthase	6	sugar metabolism
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 170 genes not included in the v1.0 annotation.

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

Table S7. Additional gene models identified in selected families.

Gene family	Gene models in V1.0 annotation	Additional gene models*	Total Brachypodium genes	Oryza	Sorghum	Arabidopsis	Populus
F-box	427	170	597	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	49	nd	80	nd

*All new models were supported by expression evidence.

S4 Prediction of the Brachypodium Secreted Proteome

A comparative survey was conducted of the predicted secreted proteome of Brachypodium, Arabidopsis and rice, to determine whether the substantial differences between grass and dicot cell wall architectures (*22*) 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 (Table S8), was selected as generating the most accurate prediction since based on the smallest proportions of apparent false positive or negative predictions following manual inspection (not shown).

Table S8. 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 (25,432)	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 (TAIR8version), and rice RAP2 were clustered using the homolog clustering algorithm TribeMCL (*23*). 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 (Figure S7).

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 Table S5). One key difference in cell wall structure of monocots and dicots is the relatively high content of pectin in the primary cell walls of dicots (*22*). 26 pectate lyase genes were identified in Arabidopsis, 29 in poplar, but only 7 in Brachypodium, 12 in rice and 10 in sorghum. Conversely, expansins, which are proteins that disrupt hydrogen bonds between cellulose microfibrils and cross-linking glycans in the plant walls and play a major role in cell-wall extension during growth (*24*), 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). This suggests either that expansins have more than one substrate or activity in grass walls, or possibly also that they have more biological functions.

Glycosyl hydrolase family 5 (GH family 5) proteins, which include endo-1,4-β-D-glucanases, endo-xylanases and other hydrolases with different substrate preferences (CAZy database) (*25*), are potential new candidates for cell wall-

degrading enzymes. We identified 10 GH5 genes in Brachypodium and 17 each in rice and sorghum belonging to three subfamilies (Sec family 515, 1219 and 2860), compared with 13 in Arabidopsis and 25 in poplar that lacked members of the Sec family 2860. This suggested that the secreted proteins belonging to Sec family 2860, not found in Arabidopsis or poplar, may contribute to the monocot-specific cell wall metabolism.

The plant secondary cell wall contains cellulose, hemicellulose and lignin (*26*). Lignin, the second most abundant natural polymers in plant cell walls after cellulose, is largely cross-linked by the cellulose/hemicellulose matrix of the secondary cell wall. Dirigent proteins, 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).

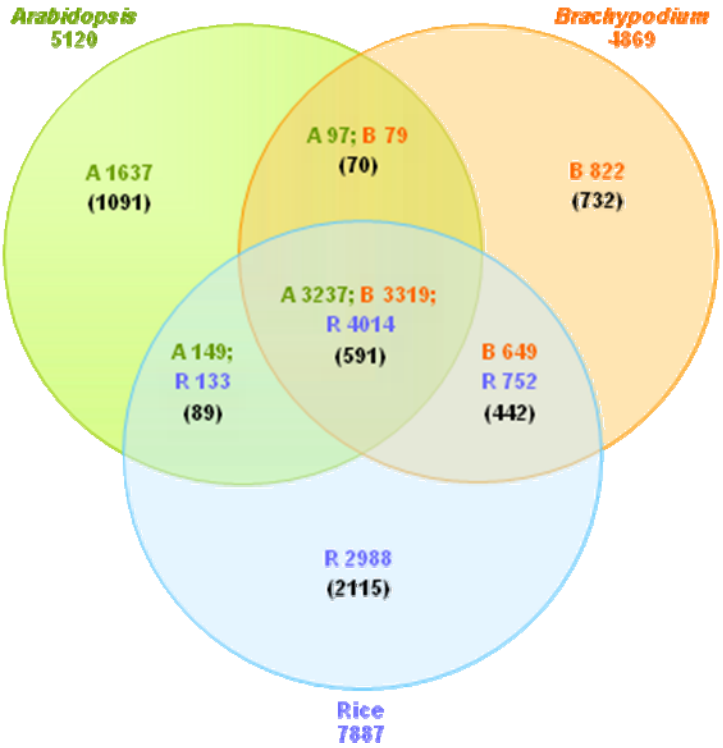


Figure S7. Venn diagram of genes carrying a predicted signal peptide between Arabidopsis (A), Brachypodium (B) and rice (R). The number of Brachypodium signal peptide-containing protein genes is similar to that of Arabidopsis. Numbers in parentheses indicate the number of ABR protein families.

Table S9. Examples of signal peptide-containing protein families from Brachypodium and rice with differential abundance in Brachypodium and Arabidopsis.

ABR 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*	*Xyloglucan fucosyltransferase (Glycosyltransferase family 37)*
	Brachypodium	*16*	*3*	
	Rice	*21*	*1*	
216	*Arabidopsis*	*26*	*23*	*Pectate lyase family protein*
	Brachypodium	*7*	*2*	
	Rice	*12*	*8*	
524	*Arabidopsis*	*17*	*17*	*Invertase/pectin methylesterase inhibitor*
	Brachypodium	*2*	*2*	
	Rice	*4*	*4*	
582	*Arabidopsis*	*1*	*1*	*Galactosyltransferase family protein (Glycosyltransferase family 31)*
	Brachypodium	*10*	*7*	
	Rice	*10*	*7*	
1029	*Arabidopsis*	*5*	*5*	*Dirigent protein family*
	Brachypodium	*0*	*0*	
	Rice	*8*	*7*	

S4. Identification of grass subfamily-specific gene sets

To identify genes and gene families that are enriched in Brachypodium and the Pooideae, Erhartoideae 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 Erhartoideae. We applied a rigorous two-way-OrthoMCL clustering schema 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 S3. A 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 have been collapsed using CD-HIT (*27*). Due to only partial representation, 3874 wheat/barley TCs/EST were not grouped with Brachypodium genes although a Brachypodium homolog was present. 16,365 Brachypodium genes clustered with representatives from wheat /barley and and 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 of cDNA 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 of all three lineages were detected. 681 families were shared between Brachypodium and rice (Erhartoideae*)* but not with sorghum, and 1,689 families were shared between Brachypodium and sorghum but not with rice. 265 families containing 811 genes and 832 singleton genes (1,643 genes; 6.54%) appeared to have only homologs in wheat and barley but not in rice or sorghum and were a potential candidate 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 potentially erroneous annotation. This further reduced the number of Pooideae- specific genes without counterparts in rice and sorghum to 1,400 (5.6%). This data is shown in Supplementary Figure S3.

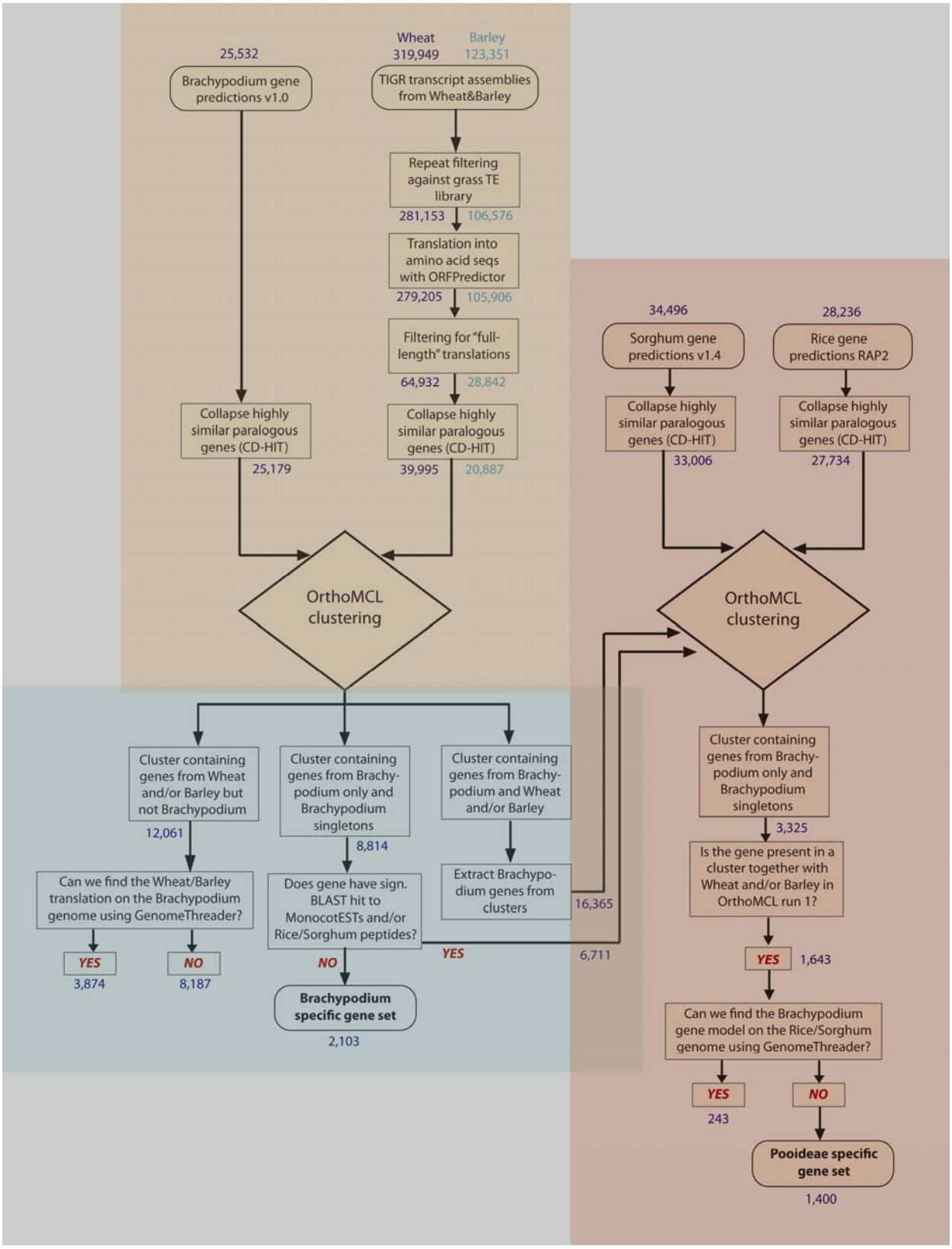


Figure S8. Workflow of two-way orthoMCL analysis to detect Brachypodium- and Pooideae-specific genes.

S5. Grass- family and species- specific gene functional categories

The blast2go suite (*28*) 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 and at least 5 distinct genes of 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 Figure 2C and Figure S8, and for tandem repeat genes described in Figure S9 below.

Table S10. 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 (4th column). Related or correlated functional categories are highlighted with the same background color, 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 harbor orthologs in Arabidopsis, Brachypodium, sorghum and rice, describing an angiosperm core set.

B. Monocot core orthologs that are shared in Brachypodium, sorghum and rice but lack a detectable ortholog in Arabidopsis.

C. the set of Pooideae specific orthologs that were obtained by the orthoMCL scheme described in Figure S3

D. Brachypodium specific genes.

S10 A. Angiosperm core 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

S10 B. Monocot-specific conserved 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

S10 C. 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

S10 D. 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

S6. Identification of tandem repeat genes

An undirected graph with genes as nodes and protein similarities as edge weights were 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%) tandemly 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.

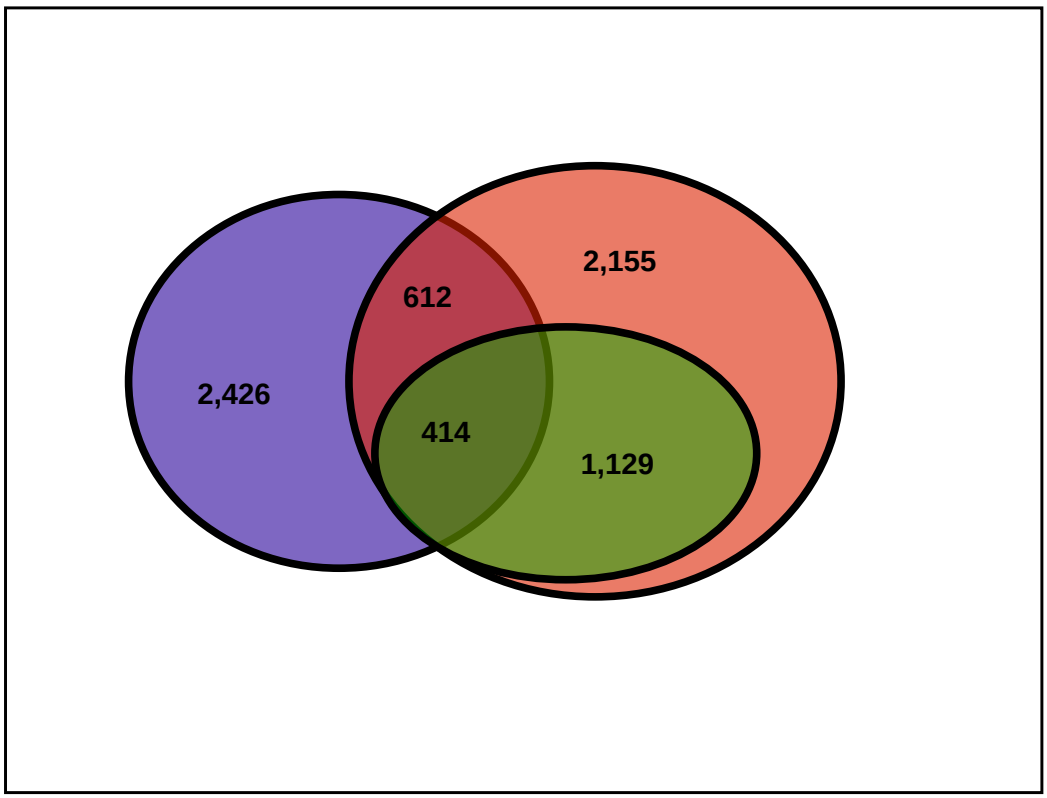


Figure S9. Tandemly repeated genes contribute disproportionately to monocot-specific gene functions.

Tandem genes (blue) comprise 3,452 loci (13.5%) out of 25,532 loci in the whole genome. 4,870 Brachypodium loci represent the grass core gene set (red) for which the four-way orthoMCL analysis detected orthologs in all three grass species but not in Arabidopsis. A total of 1,026 of these represent tandem genes. Out of 4,870 monocotyledonous core genes, 1,543 were associated with significantly enriched functional categories. 414 (26.8%) of these genes were tandemly repeated genes . ($P<10^{-16}$, fisher's exact test).

Table S11. Gene functions enriched in tandem repeat genes

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 o
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
GO:0016798	408	100	3.591395e-002	hydrolase activity, acting on glycosyl bonds
GO:0000166	3521	676	4.031522e-002	nucleotide binding

S7. Repeats Analysis

LTR retrotransposons

De novo searches for LTR retrotransposons were performed with LTR_STRUCT and LTR_HARVEST (*29*). 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. 693 (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 intact 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 over 16.1% of the genome by nucleotides, 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 between chromosomes 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%, even though chromosome 5 is only 58% the size of chromosome 4. Chromosome 4 is deficient in Gypsy elements (2.34 X more abundant), whereas chromosome 5 is enriched (2.9 X). 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 where added to mipsREdat (mips.gsf.de/proj/plant/webapp/recat/), a plant repeat element database, and used for the 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 (*30*). 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 (*31*).

A total of 1814 solo LTRs was identified in Brachypodium by similarity search to the full-length elements and by structural analysis, representing only 0.25% of the current genome size. However, assuming that each one (average length 379 bp) was derived from an intact element of 10 kb, a minimum of 17.4 Mb has 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 (1122) are 1.6-fold more abundant than the *Copia* solo LTRs (689), similar to the relative abundance of intact *Gypsy* elements (1.36). The solo LTRs are on average 4.3 MY old (*Gypsy* 4.32 MY, *Copia* 4.35 MY), based on the sequence divergence time from the most similar intact element. The youngest solo LTR is 58 thousand years old, and

the old 11.7 MY (Figure S4). Of all the intact elements in the *B. distachyon* 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. Neither Bd3_RLG_17 has not been previously annotated, but is 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 20769 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 20000 years old, and has two solo LTRs. of retrotransposons in *B. distachyon.* Given that the *Angela – BARE – Wis* family members are among the recently active members of the *B. distachyon* genome, this is further evidence for the role of retrotransposon loss through recombination as a way of controlling genome size expansion.

Differences between the chromosomes concerning solo LTR distribution are striking. While the chromosomes have on average 362 solo LTRs each, chromosome 5 is notably poorest, with only 73 in total, whereas chromosome 3 has 1016. 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, those matching 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 repetitious 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 intact elements, chromosome 5 has a ratio of only 0.89, in contrast to chromosome 3 with 6.96; the others have ratios between 1.23 and 1.73. These data, taken together with those on the number and age of the full-length LTR retrotransposons, suggests that chromosome 5 is gaining retrotransposons by replication and losing comparatively few by recombination.

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 by a priority based approach. High confidence expert annotations are assigned first, with a higher priority on young full length elements, which still posses target site duplications. Overlapping elements with lower priority are either truncated, fragmented or skipped, depending on adjustable parameters for overlap percent and minimum rest length. The assignment order within one priority group is defined by descending homology score or element length. For Brachypodium all elements overlapping > 80% of their length to higher priority elements were removed. Elements overlapping <=80% where truncated or split, if the remaining length exceeded 49 bp. In a first step overlaps within each of the 10 different annotations where removed. The following priority order was used in 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-retrotransposons fragments. Step 1-7 where applied in 2 iterations, first with full length elements still having target site duplications, 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 Table 2.

Simple Sequence Repeats

For SSRs searches SSRLocator (*32*) was used. It was configured to locate perfect, imperfect and composite SSRs (*33*) Class I (≥ 20 bp) and Class II (≥ 12 and < 20 bp) repeats (*34*), which correspond to 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 (*35, 36*). 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 Table S12. 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%) are assessed. G/C-rich trimers, independent from sequence arrangement motifs, predominate (35%). For tetramer, pentamer and hexamer motifs, no apparent predominance of a given motif was detected. SSRs are overwhelmingly present in intergenic (88.0%) regions 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.

Table S12. Summary of simple sequence repeat (SSR) types and numbers in the Brachypodium genome.

Type	Class	Total Loci	Total Repeats (nº repeats)	Total Length (bp) (nº repeats * type)	Average Length (bp) (Total length / Total loci)	Repeat Numbers
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**	

Occurrence	Repeat Total	%	bp total	%	Average Number repeats	ssr/mb
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 (*37*) (*37*)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 (*38*) (*38*)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 (*39*). A spike sequence is used to reduce the noise in the BLAST results (*40*).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. We identified 18,664 putative conserved non-coding sequences in the Brachypodium genome with lengths ranging from 4 to 2255 nucleotides (Figure S6: mean length 28 bp, median length 21 bp, 0.87 CNS per gene) using these stringent criteria. The majority of Brachypodium genes have no CNS, 4008 genes have one CNS and 153 genes have more than 10 CNS each. In order to determine whether the identified CNS contained potentially functional motifs we took a set of 392 rice genes shown experimentally to be up-regulated in drought conditions (*41*) and identified 321 orthologs in Brachypodium using BLAST and an e-value cutoff of e-50. We identified 357 associated CNS in which conserved DRE/CRT drought response motifs (*42*) were significantly over-represented (χ^2 (1, N=43759) = 4.57, $p<0.05$). An example of a CNS containing a DRE/CRT cis-acting element is shown in Figure S8.

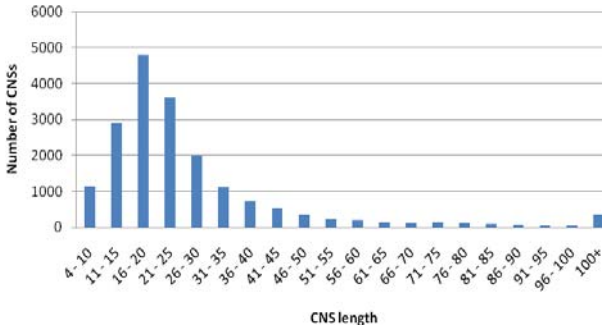


Figure S10. Distribution of CNS lengths

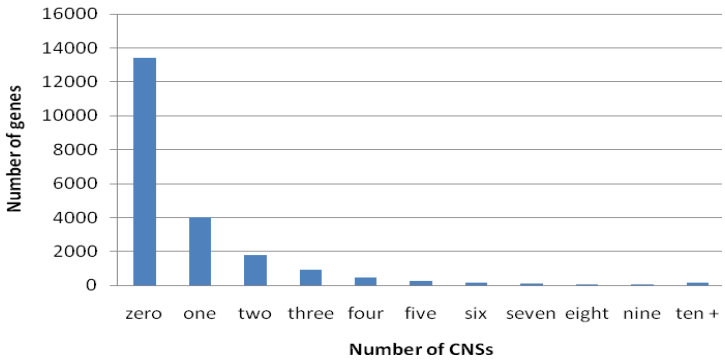


Figure S11. Distribution of the number of CNS per gene

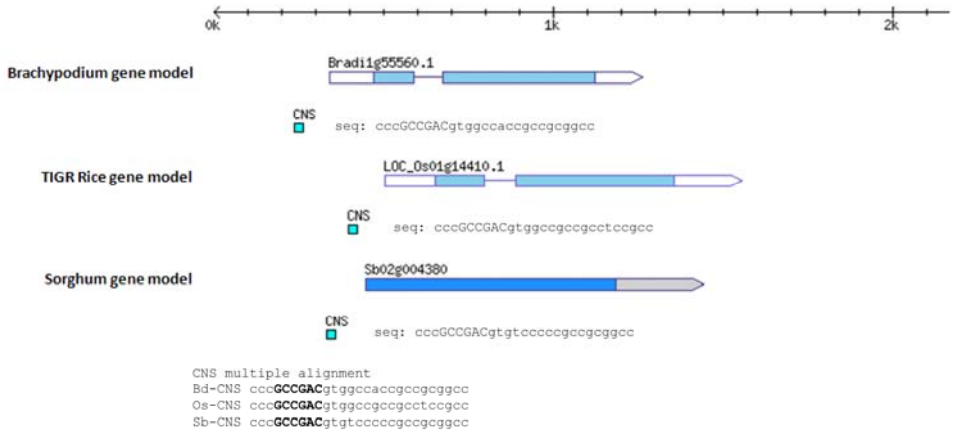


Figure S12. A conserved non-coding sequence element upstream of orthologous genes in Brachypodium, rice and sorghum. The multiple sequence alignment shows the core DRE/CRT (dehydration-responsive element/C-repeat) cis-acting element in bold.

S 8. Ks analysis of whole genome ortholog comparisons

Orthologous genes of Brachypodium were determined in rice (TIGR5) and sorghum (v1.4) genes as described in S4 previously. 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) (*43*) were generated for each orthologous protein pair and transformed to pairwise codon based alignments. Codeml of the PAML package (*44*) using the F3x4 model was applied to estimate Ka and Ks by maximum-likelihood and by the method of (*45*).

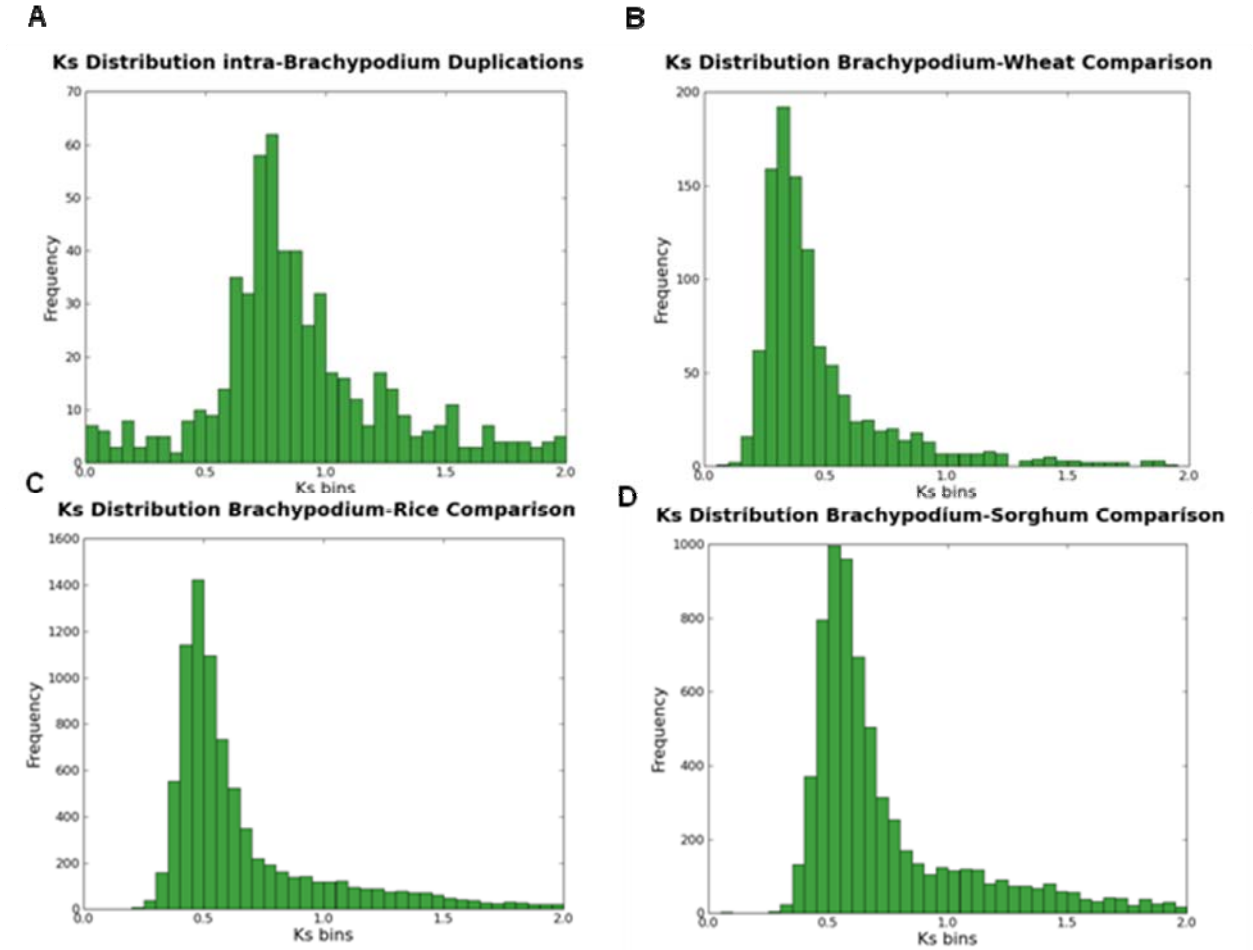


Figure S13. Ks Distributions of intra-genomic Brachypodium duplications and Brachypodium- sorghum-rice-wheat and -maize orthologous genes.

The charts show Ks values derived by the maximum-likelihood method (*44*). The bin size of Ks values is 0.05. Note that the maize and 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.

Table S13. 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 $\lambda=6.1\times10^{-9}$ (mean of $5.1\text{-}7.1\times10^{-9}$) (*46*). Ks estimates for wheat may be overestimated as they are based on EST data. Figure 4B 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

S9. 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 (*47*) 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 (*48*), and genetic maps of 1,015 barley ESTs (*49*) and 863 *Ae. tauschii* ESTs (*50*). Gene relationships and order were compared using the CIP-CALP method (*48*). 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).

Table S14. 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.

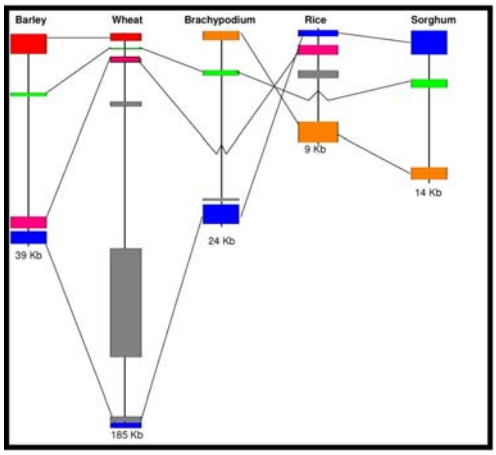
Internode	Time * (MY)	Inversions and subchrom. translocations(No.)	Rate No. changes MY^{-1}	All changes (No.)	Rate No. changes MY^{-1}
Brachypodium	29.4	5	0.17	12	0.41
Ae. tauschii	29.4	36	1.22	41	1.39
Brachypodium + *Ae.t.*	12.2	1	0.08	1	0.08
Rice	42.1	4	0.10	4	0.10
Sorghum	50.5	5	0.10	2	0.14
Could not be assigned		7		7	

* For time estimates see Figure 4.

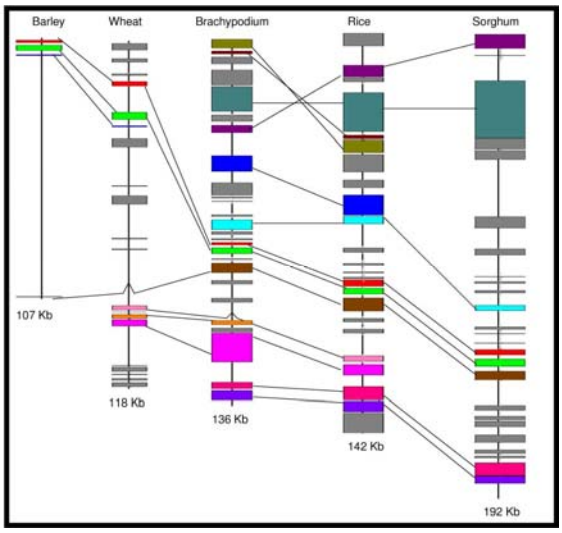
The linear order of 863 gene loci mapped on the *Ae. tauschii* EST genetic map (*50*) and orthologous loci in Brachypodium, rice and sorghum were used to estimate the rates of chromosome evolution at the internodes of their phylogenetic tree (Fig. 4B). The following strategy was used to assign changes 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 Ehrhratoideae (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 4A. 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 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 subchromosome-size changes, 14 chromosome-size translocations resulting in the dysploid reductions of the basic chromosome number were assigned to three terminal internodes (Table S9). 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.

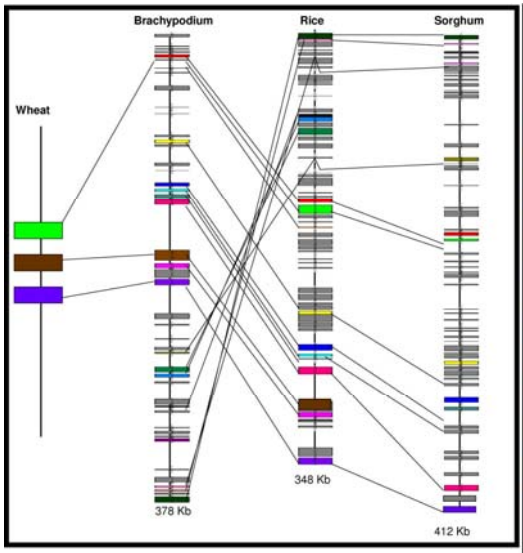
***Glu* locus**



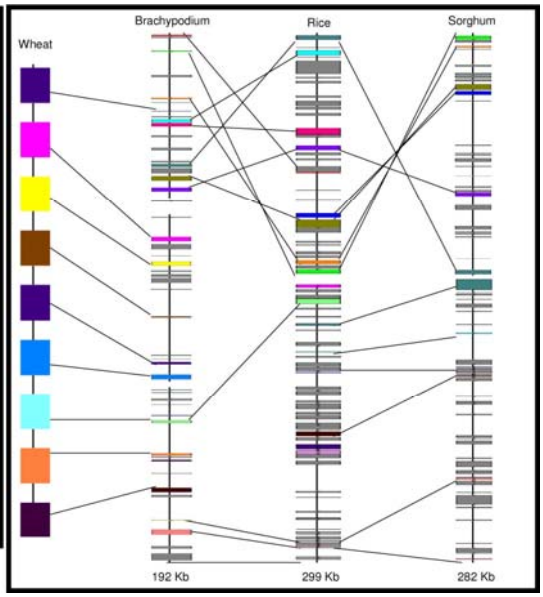
***Ha* locus**



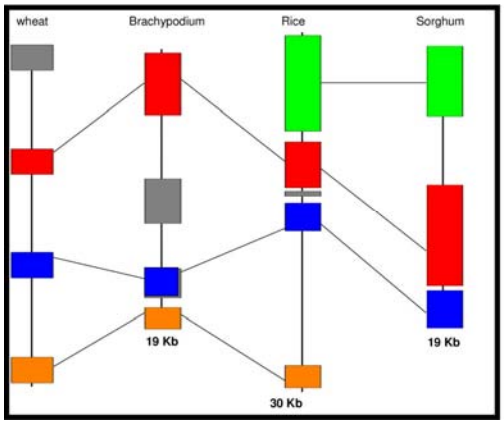
***Lr34* locus**



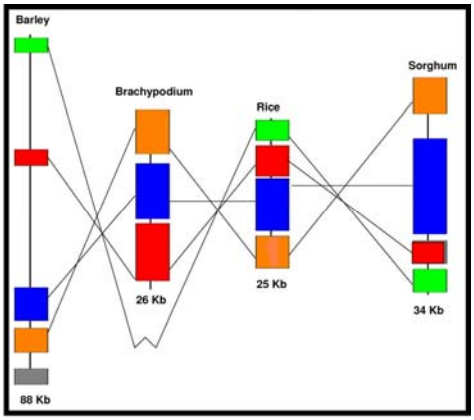
***Ph1* locus**



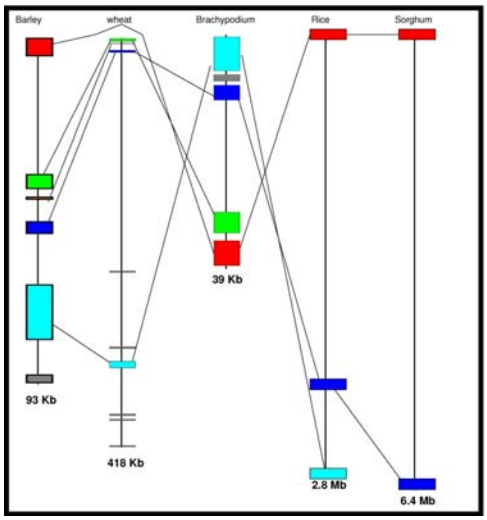
***Sh2/a1* locus**



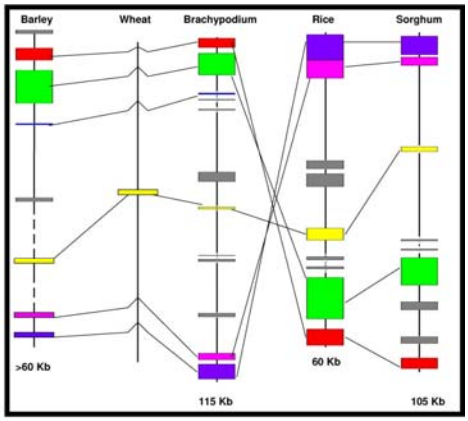
***Vrn1* locus**



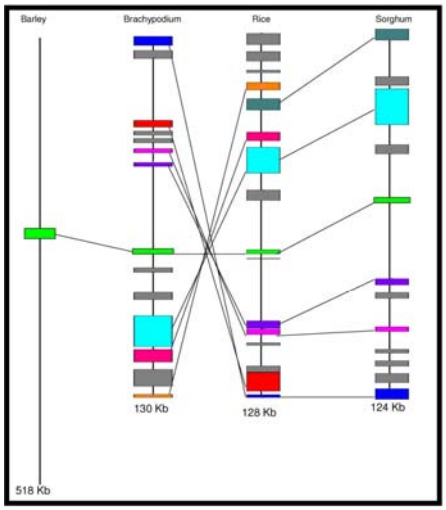
***Vrn2* locus**



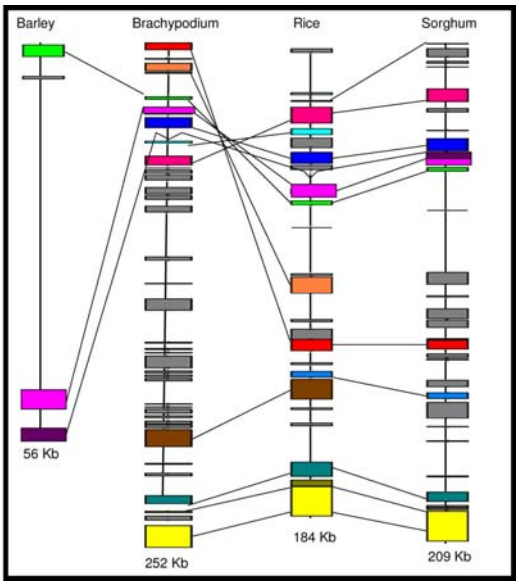
***Vrn3* locus**



***Vsr1* locus**



***BCD135* locus**



***Ppd-H1* locus** ***Rph7* locus**

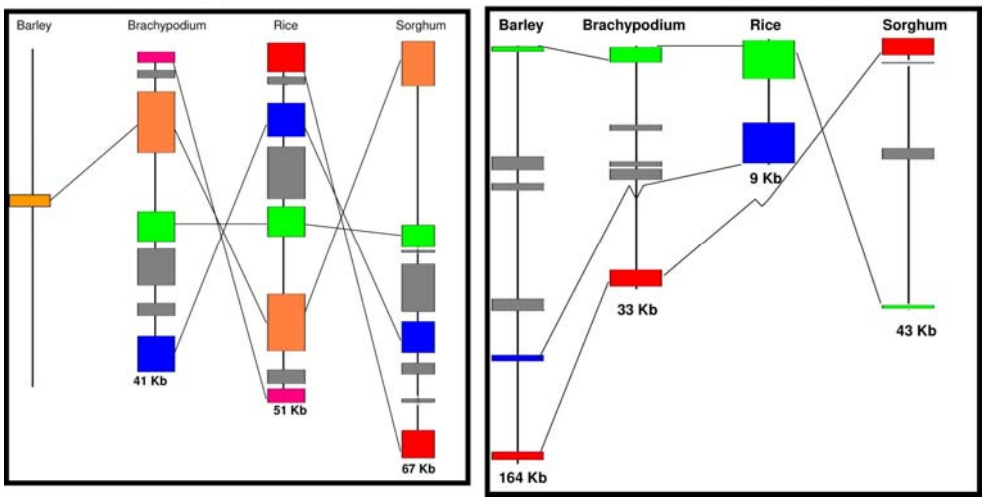


Figure S14. Microsynteny analysis between rice, sorghum and Brachypodium at the Triticeae *Ha*, *Glu*, *Lr34*, *Ph1*, *Sh2/a1*, *Vrn1*, *Vrn2*, *Vrn3*, *Vrs1*, *BCD135*, *Phd-H1,* Rph7 loci. Annotated genes are illustrated with squares and collinear genes are illustrated with the same color code. Micro-collinearity analysis at 12 specific loci for which wheat or barley BAC sequences (covering a total 1.9 Mb) are available (*Ha*, (*51*); *Glu*, (*52*); *Lr34*, (*53*); *Ph1*, (*54*); *Sh2/a1*, (*55*); *Vrn1*, (*56*); *Vrn2*, (*57*); *Vrn3*, (*58*); *Vrs1*, (*59*); *BCD135*, (*60*); *Ppd-H1*, (*61*); Rph7 (*62*). This demonstrated that at diverse loci 62.5% of genes are conserved between the Triticeae and Brachypodium, compared to less that 55% between the Triticeae, sorghum and rice.

Table S15. Large Brachypodium gene families and their degree of collinearity in rice and sorghum.

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%
Fbox	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.

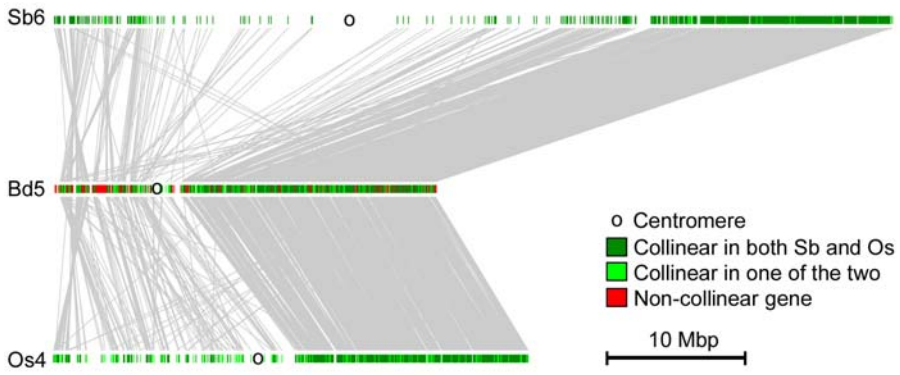


Figure S15. Map of Brachypodium chromosome 5 (Bd5) and its syntenic chromosomes from sorghum (Sb6) and rice (Os4). Collinear genes are connected by grey lines. In all three species the short arm has lower gene density, reduced collinearity and multiple rearrangements such as inversions and translocations.

S10. Small RNA library construction and sequencing.

Brachypodium Bd21 was used for the preparation of two panicle (flower) libraries. For OBD01, plants were grown in long-day conditions (16 h days/8 h nights) at 25°C. Inflorescence tissue was collected (day 28-35) at 4 time point intervals of 0700 (dawn), 1300, 1900, 0100 hours, and frozen immediately on 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 BD105 was approximately 120-140 umol m^{-2} sec^{-1}. OBD01 total RNA was extracted using Trizol reagent (Invitrogen) as described in (*63*) 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 (*64*) 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 comigrated with 32P-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 BD105 panicle tissues, total RNA was isolated using Trizol reagent and small RNA libraries were constructed according to (*65, 66*). 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 (*67*), 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 = $\ln\left[\left(1 + 10 \times \frac{\sum_{i=1}^{10} P_i}{1 + \sum U}\right)^{n-2}\right]$, 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 (*67*). 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 *B. distachyon* 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, wildtype Arabidopsis inflorescence library available in GenBank's GEO as GSM284747.

BDI05

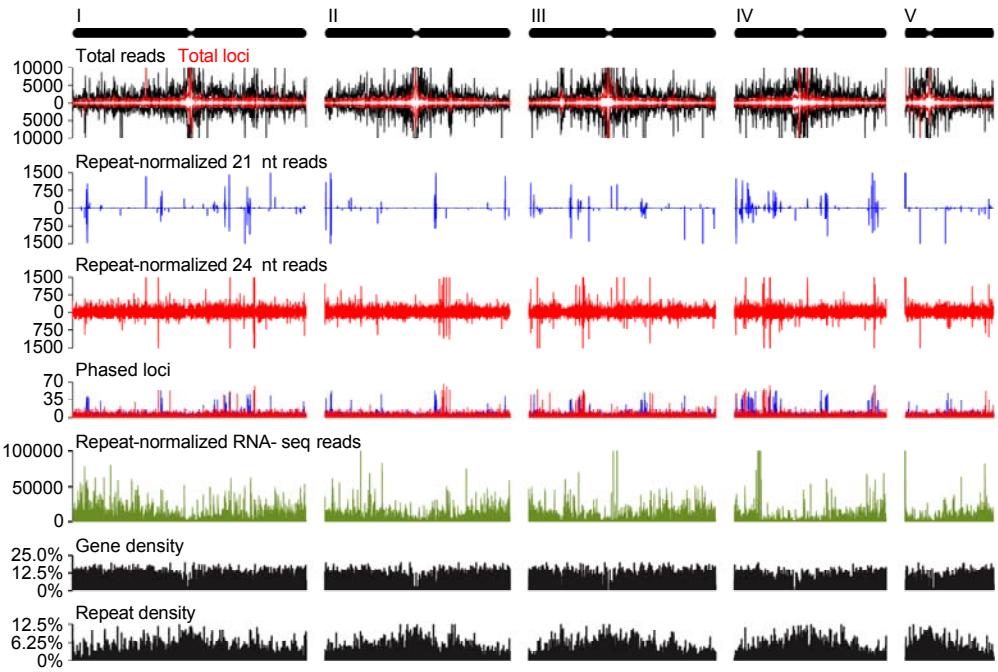


Figure S16. Genome-wide distribution of small RNA genes identified in the BD105 panicle library and their alignment with repeat elements in the Brachypodium genome. Each of the five Brachypodium chromosomes are shown as ideograms at the top of each figure. Total reads and total loci graphs plot total 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 = 100000 nt, scroll = 20000 nt).

Table S16. Scores for analysis of small RNA phasing intervals in the *B. distachyon* genome.

	Interval[a] →	19		20		21		22		23		24		25	
	Phasing score ↓	position number	cluster number	position number	cluster number	position number	cluster number	position number	cluster number	position number	cluster number	position number	cluster number	position number	cluster number
Arabidopsis inflorescence[b]	>7.5	29,113	2,601	22,985	2,295	18,696	2,082	14,607	1,786	12,049	1,661	10,386	1,545	9,386	1,398
	>10	3,640	792	2,962	679	2,343	537	1,696	426	1,251	342	1,118	330	918	278
	>12.5	384	112	416	118	401	91	260	78	175	35	153	36	132	46
	>15	94	14	75	19	182	26	66	7	84	4	73	5	49	12
	>17.5	36	2	33	4	100	17	39	5	43	2	26	3	18	5
	>20	13	2	13	3	53	13	12	2	14	3	7	3	4	1
	>22.5	7	2	5	2	29	7	0	0	2	2	0	0	0	0
	>25	0	0	0	0	21	5	0	0	0	0	0	0	0	0
	>27.5	0	0	0	0	7	4	0	0	0	0	0	0	0	0
	>30	0	0	0	0	2	2	0	0	0	0	0	0	0	0
B. distachyon panicles BDI05	>7.5	11,269	3,365	10,177	3,073	16,421	3,517	7,399	2,392	6,537	2,160	11,196	2,254	5,327	1,766
	>10	2,920	819	2,616	750	9,085	1,551	1,749	538	1,566	452	6,217	748	1,399	398
	>12.5	854	211	801	201	6,587	1,074	497	144	449	113	4,635	393	516	120
	>15	306	67	271	65	5,140	838	160	43	135	45	3,882	299	189	46
	>17.5	113	28	81	25	4,083	693	51	18	32	10	3,414	257	30	15
	>20	50	14	17	8	3,224	589	18	10	2	2	3,056	227	10	5
	>22.5	11	7	6	5	2,519	509	2	2	0	0	2,756	213	0	0
	>25	1	1	2	2	1,865	413	1	1	0	0	2,462	198	0	0
	>27.5	0	0	0	0	1,348	329	1	1	0	0	2,203	188	0	0
	>30	0	0	0	0	951	252	0	0	0	0	1,924	180	0	0
B. distachyon panicles OBD01	>7.5	13,467	2,917	11,767	2,671	18,887	3,302	8,661	2,209	7,625	1,906	11,787	2,077	6,094	1,537
	>10	3,425	805	2,846	708	10,410	1,592	1,852	558	1,701	487	5,893	749	1,435	358
	>12.5	769	232	683	205	7,687	1,146	384	144	377	134	4,353	409	325	96
	>15	190	62	173	61	6,387	986	118	56	132	36	3,750	303	114	34
	>17.5	55	19	55	26	5,355	877	43	20	59	20	3,404	254	61	11
	>20	13	9	24	14	4,472	776	17	12	32	12	3,088	235	40	8
	>22.5	2	2	4	3	3,625	668	10	7	24	7	2,797	227	31	6
	>25	1	1	0	0	2,876	579	4	2	14	5	2,504	217	16	6
	>27.5	0	0	0	0	2,236	473	1	1	11	5	2,240	210	9	5
	>30	0	0	0	0	1,661	386	0	0	8	5	1,976	190	6	5

Gray regions of table indicate small RNAs or clusters of particular interest, exceeding an arbitrary cut-off score of 25. “Position number” indicates the number of sites matched by small RNAs that had at or above a specific score, “cluster number” indicates the number of loci at or above the score; all high scoring positions within a 300 bp window were combined to generate one cluster.

[a] Interval indicates the number of nucleotides between small RNAs, analyzed in a 10-phase window across the genome. The algorithm is described in more detail in the Supplemental Methods section.

[b] The Arabidopsis small RNA library was previously described (*68*).

S11. *In situ* hybridization

Metaphase chromosome spreads were made from excised and fixed Brachypodium Bd21 roots grown for 3-5 days, essential as described (*69*). BACs were identified for labelling from a physical map of Brachypodium (*5*) that is integrated with genome sequence assemblies. Reference BACs with known chromosomal locations (*6*) were 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* (*70*) was used to visualise the 45S rDNA locus that is diagnostic for short arm of chromosome 5. A 5S rDNA probe was obtained from the wheat clone pTa794 (*71*) 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) hybridisation mixture consisted *inter alia* of 50% deionised formamide, 20% dextran sulphate, 2× SSC and salmon sperm blocking DNA in 25-100× excess of labelled probes. All probes were mixed to a final concentration each of 2 - 5 ng/μl of the mixture and pre-denatured (75 ºC for 10 min). The slides with chromosome material and the hybridisation 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 carries out for 10 min in 10% deionised formamide in 0.1× SSC at 42 ºC, which provides the stringency allowing to leave DNA-DNA hybrids with a sequence identity of 79%. All digoxigenated probes were immunodetected using standard protocol for FITC-conjugated anti-digoxigenin antibodies (Roche) and were visualised 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).

S12 References

1. D. G. Peterson *et al.*, *Plant Molecular Biology Reporter* **15**, 148 (1997).
2. N. Huo *et al.*, *Genome* **49**, 1099 (2006).
3. N. Huo *et al.*, *Funct Integr Genomics* **8**, 135 (2008).
4. Y. O. Gu *et al.*, *BMC Genomics in press* (2009).
5. M. Febrer *et al.*, *BMC Genomics in press* (2009).
6. R. Hasterok *et al.*, *Genetics* **173**, 349 (2006).
7. G. Gremme *et al.,Information and Software Technology* **47**, 965 (2005).
8. S. Ouyang *et al.*, *Nucleic Acids Res* **35**, D883 (2007).
9. T. Tanaka *et al.*, *Nucleic Acids Res* **36**, D1028 (2008).
10. A. H. Paterson *et al.*, *Nature* **457**, 551 (2009).
11. Arabidopsis Genome Initiative *Nature* **408**, 796 (2000).
12. S. Y. Rhee *et al.*, *Nucleic Acids Res* **31**, 224 (2003).
13. J. E. Allen, S. L. Salzberg, *Bioinformatics* **21**, 3596 (2005).
14. K. Mayer *et al.*, *Nature* **402**, 769 (1999).
15. T. M. Lowe, S. R. Eddy, *Nucleic Acids Res* **25**, 955 (1997).
16. S. Fox, S. Filichkin, T. Mockler, Eds., *Applications of ultra high throughput sequencing*, vol. 553 (Humana Press, 2009).
17. T. Imaizumi, S. A. Kay, *Trends Plant Sci* **11**, 550 (2006).
18. J. Colasanti, V. Coneva, *Plant Physiol* **149**, 56 (2009).
19. C. Dardick, P. Ronald, *PLoS Pathog* **2**, e2 (2006).
20. J. M. Escobar-Restrepo *et al.*, *Science* **317**, 656 (2007).
21. K. Hematy, H. Hofte, *Curr Opin Plant Biol* **11**, 321 (2008).
22. J. Vogel, *Curr Opin Plant Biol* **11**, 301 (2008).
23. A. J. Enright *et al.*, *Nucleic Acids Res* **31**, 4632 (2003).
24. D. J. Cosgrove, *Nature* **407**, 321 (2000).
25. B. L. Cantarel *et al.*, *Nucleic Acids Res* **37**, D233 (2009).
26. M. B. Sticklen, *Nat Rev Genet* **9**, 433 (2008).
27. W. Li, A. Godzik, *Bioinformatics* **22**, 1658 (2006).
28. S. Gotz *et al.*, *Nucleic Acids Res* **36**, 3420 (2008).
29. E. M. McCarthy, J. F. McDonald, *Bioinformatics* **19**, 362 (2003).
30. J. Ma, J. L. Bennetzen, *Proc Natl Acad Sci U S A* **101**, 12404 (2004).
31. T. Wicker, B. Keller, *Genome Res* **17**, 1072 (2007).
32. L. C. da Maia *et al.*, *Int J Plant Genomics* **2008**, 412696 (2008).

33. R. K. Varshney, T. Thiel, N. Stein, P. Langridge, A. Graner, *Cell Mol Biol Lett* **7**, 537 (2002).
34. S. Temnykh *et al.*, *Genome Res* **11**, 1441 (2001).
35. D. Field, C. Wills, *Proc Biol Sci* **263**, 209 (996).
36. J. Jurka, C. Pethiyagoda, *J Mol Evol* **40**, 120 (1995).
37. L. Li, C. J. Stoeckert, Jr., D. S. Roos, *Genome Res* **13**, 2178 (2003).
38. T. A. Tatusova, T. L. Madden, *FEMS Microbiol Lett* **174**, 247 (1999).
39. N. J. Kaplinsky *et al.*, *Proc Natl Acad Sci U S A* **99**, 6147 (2002).
40. E. Lyons, M. Freeling, *Plant J* **53**, 661 (2008).
41. T. Degenkolbe *et al.*, *Plant Mol Biol* **69**, 133 (2009).
42. Q. Liu *et al.*, *Plant Cell* **10**, 1391 (1998).
43. P. Rice *et al.*,*Trends Genet* **16**, 276 (2000).
44. Z. Yang, *Mol Biol Evol* **24**, 1586 (2007).
45. M. Nei, T. Gojobori, *Mol Biol Evol* **3**, 418 (1986).
46. K. H. Wolfe, P. M. Sharpe, W. H. Li, *J Mol Evol* **29**, 208 (1989).
47. L. L. Qi *et al.*, *Genetics* **168**, 701 (2004).
48. J. Salse *et al.*, *Plant Cell* **20**, 11 (2008).
49. N. Stein *et al.*, *Theor Appl Genet* **114**, 823 (2007).
50. M. C. Luo *et al.*, *Proc. Natnl. Acad. Sci. U.S.A. in press* (2009).
51. N. Chantret *et al.*, *Plant Cell* **17**, 1033 (2005).
52. X. Y. Gu *et al.*,*Genetics* **166**, 1503 (2004).
53. E. Bossolini *et al.*,*Plant J* **49**, 704 (2007).
54. S. Griffiths *et al.*, *Nature* **439**, 749 (2006).
55. J. L. Bennetzen, J. Ma, *Curr Opin Plant Biol* **6**, 128 (2003).
56. W. Ramakrishna *et al.*, *Genetics* **162**, 1389 (2002).
57. L. Yan *et al.*, *Science* **303**, 1640 (2004).
58. L. Yan *et al.*, *Proc Natl Acad Sci U S A* **103**, 19581 (2006).
59. M. Pourkheirandish *et al.*,*Theor Appl Genet* **114**, 1357 (2007).
60. Y. Park, *et al., Mol Cells* **17**, 492 (2004).
61. R. P. Dunford *et al.*, *Genetics* **161**, 825 (2002).
62. S. Brunner *et al.*,*Genetics* **164**, 673 (2003).
63. C. Llave *et al., Plant Cell* **14**, 1605 (2002).
64. K. D. Kasschau *et al.*, *PLoS Biol* **5**, e57 (2007).
65. C. Lu *et al.*, *Science* **309**, 1567 (2005).
66. C. Lu *et al.*,*Methods* **43**, 110 (2007).
67. M. D. Howell *et al.*, *Plant Cell* **19**, 926 (2007).
68. R. Lister *et al.*, *Cell* **133**, 523 (2008).
69. G. Jenkins, R. Hasterok, *Nat Protoc* **2**, 88 (2007).
70. I. Unfried, P. Gruendler, *Nucleic Acids Res* **18**, 4011 (1990).
71. W. L. Gerlach, T. A. Dyer, *Nucleic Acids Res* **8**, 4851 (980).