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

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

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genetic markers, a genetic linkage map, bacterial artificial chromosome (BAC) libraries, physical maps (M.F., unpublished observations), mutant collections (http://brachypodium.pw.usda.gov, http://www.brachytag.org), microarrays and databases (http://www.brachybase.org,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

**Genome sequence assembly and annotation**

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

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

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

Between 77 and 84% of gene families (defined according to Supplementary Fig. 6) are shared among the three grass subfamilies represented by *Brachypodium*, rice and sorghum, reflecting a relatively
The abundance and distribution of the following genome elements have different ranges and different maximum (max) pseudocolour levels: STA cent base-pair (%bp) coverage of the respective window. The heat map tracks gene introns and satellite tandem arrays (STA). Graphs are from 0 to 100 per repeat transposable elements (MITEs); MITEs; gene exons (CDS); potentially autonomous DNA transposons that are not miniature inverted-repet transposable elements (MITEs) (DNA-TEs); MITEs; gene exons (CDS); DNA transposons (LTR); solo-LTRs (sLTRs); complete LTR retroelements (cLTRs); retrotransposons.

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

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

models required modification and very few pseudogenes were identified, demonstrating the accuracy of the v1.0 annotation. Phylogenetic trees for 62 gene families were constructed using genes from rice, Arabidopsis, sorghum and poplar. In nearly all cases, Brachypodium genes had a similar distribution to rice and sorghum, demonstrating that Brachypodium is suitably generic for grass functional genomics research (Supplementary Figs 8 and 9). Analysis of the predicted secretome identified substantial differences in the distribution of cell wall metabolism genes between dicots and grasses (Supplementary Tables 12, 13 and Supplementary Fig. 10), consistent with their different cell walls26. Signal peptide probability curves also suggested that start codons were accurately predicted (Supplementary Fig. 11).

Maintaining a small grass genome size

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

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

Whole-genome comparison of three diverse grass genomes

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

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

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

©2010 Macmillan Publishers Limited. All rights reserved
Bd5S, Os4S and Sb6S also have the loweststrate maintenance of this high repeat content forSyntenic regions of rice (Os4S) and sorghum (Sb6S) demon-
transposons by replication and losing fewer by recombination.
unlike the rest of theyoungest intacthalf of the rest of the genome, high LTR retrotransposon density, the
rate of genome change in
sorghum, rice andmaps. Nested insertions were not identified in other chromosomes, possibly
chromosome insertions in barley (H chromosomes) inferred from genetic
consistent with the rapid diversification of the NBS-LRR and
(NBS-LRR and F-box) were almost never found in syntenic order
relatively large gene families, four were highly syntenic and two
a
centromeric regions explains the observed syntenic break points. Bd5 has
point in which high retrotransposon density had deleterious effects
propose that the chromosome ancestral to Bd5S reached a tipping
proportion of collinear genes (Fig. 4a and Supplementary Fig. 19). We
766
As the first genome sequence of a pooid grass, the
ARTICLES
NATURE| Vol 463| 11 February 2010
Grass crop improvement for sustainable fuel44 and food45 produc-
tion requires a substantial increase in research in species such as
Miscanthus, switchgrass, wheat and cool season forage grasses. These
considerations have led to the rapid adoption of Brachypodium as an
experimental system for grass research. The similarities in gene content
and gene family structure between Brachypodium, rice and sorghum support the value of Brachypodium as a functional genomics model
for all grasses. The Brachypodium genome sequence analysis reported
here is therefore an important advance towards securing sustainable supplies of food, feed and fuel from new generations of grass crops.

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

Comparisons of evolutionary rates between Brachypodium, sorghum, rice and Ae. tauschii demonstrated a substantially higher rate of genome change in Ae. tauschii (Supplementary Table 17). This may be due to retroelement activity that increases syntenic disruptions, as proposed for chromosome 5S later41. Among seven relatively large gene families, four were highly syntenic and two (NBS-LRR and F-box) were almost never found in syntenic order when compared to rice and sorghum (Supplementary Table 18), consistent with the rapid diversification of the NBS-LRR and F-box gene families42.

The short arm of chromosome 5 (Bd5S) has a gene density roughly half of the rest of the genome, high LTR retrotransposon density, the youngest intact Gypsy elements and the lowest solo LTR density. Thus, unlike the rest of the Brachypodium genome, Bd5S is gaining retrotransposons by replication and losing fewer by recombination. Syntenic regions of rice (Os4S) and sorghum (Sb6S) demonstrate maintenance of this high repeat content for ~50–70 Myr (Supplementary Fig. 19)41. Bd5S, Os4S and Sb6S also have the lowest proportion of collinear genes (Fig. 4a and Supplementary Fig. 19). We propose that the chromosome ancestral to Bd5S reached a tipping point in which high retrotransposon density had deleterious effects on genes.

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

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

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

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

Received 29 August; accepted 9 December 2009.


**Supplementary Information**

The whole-genome shotgun sequence of *Brachypodium distachyon* has been deposited at DDBJ/EMBL/GenBank under the accession ADDN00000000. (The version described in this manuscript is the first version, accession ADDN01000000.) EST sequences have been deposited with dbEST (accessions 67946317–68008384).

Acknowledgements

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

**Author Information**

The whole-genome shotgun sequence of *Brachypodium distachyon* has been deposited at DDBJ/EMBL/GenBank under the accession ADDN00000000. (The version described in this manuscript is the first version, accession ADDN01000000.) EST sequences have been deposited with dbEST (accessions 67946317–68008384).

**Author Contributions**

See list of consortium authors below.

**The International Brachypodium Initiative**

**Principal investigators**

John P. Vogel1, David F. Garvin2, Todd C. Mockler3, Jeremy Schmutz4, Dan Rokhsar5, Michael W. Bevan6,7; *DNA sequencing and assembly* and *Kerry Barby*8, Susan Lucas8, Miranda Harmon-Smith9, Kathleen Lai9, Hope Tice9, Jeremy Schmutz4 (Leader), Jane Grimwood10, Neil McKenzie11, Michael W. Bevan6,7.

**Pseudomolecule assembly and BAC end sequencing**


©2010 Macmillan Publishers Limited. All rights reserved

767