Funct Integr Genomics
DOI 10.1007/s10142-007-0062-7

ORIGINAL PAPER

The nuclear genome of *Brachypodium distachyon*: analysis of BAC end sequences

Naxin Huo · Gerard R. Lazo · John P. Vogel · Frank M. You · Yaqin Ma · Daniel M. Hayden · Devin Coleman-Derr · Theresa A. Hill · Jan Dvorak · Olin D. Anderson · Ming-Cheng Luo · Yong Q. Gu

Received: 27 July 2007 / Revised: 4 October 2007 / Accepted: 6 October 2007

Abstract Due in part to its small genome (~350 Mb), *Brachypodium distachyon* is emerging as a model system for temperate grasses, including important crops like wheat and barley. We present the analysis of 10.9% of the *Brachypodium* genome based on 64,696 bacterial artificial chromosome (BAC) end sequences (BES). Analysis of repeat DNA content in BES revealed that approximately 11.0% of the genome consists of known repetitive DNA. The vast majority of the *Brachypodium* repetitive elements are LTR retrotransposons. While *Bare-1* retrotransposons are common to wheat and barley, *Brachypodium* repetitive element sequence-1 (*BRES-1*), closely related to *Bare-1*, is also abundant in *Brachypodium*. Moreover, unique *Brachypodium* repetitive element sequences identified constitute approximately 7.4% of its genome. Simple sequence repeats from BES were analyzed, and flanking primer sequences for SSR detection potentially useful for genetic mapping are available at http://brachypodium.pw.usda.gov. Sequence analyses of BES indicated that approximately 21.2% of the *Brachypodium* genome represents coding sequence. Furthermore, *Brachypodium* BES have more significant matches to ESTs from wheat than rice or maize, although these species have similar sizes of EST collections. A phylogenetic analysis based on 335 sequences shared among seven grass species further revealed a closer relationship between *Brachypodium* and Triticeae than *Brachypodium* and rice or maize.

Keyword *Brachypodium* · BAC · Genome · Retrotransposons · Phylogeny · SSR

N. Huo and G.R. Lazo contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s10142-007-0062-7) contains supplementary material, which is available to authorized users.

N. Huo · G. R. Lazo · J. P. Vogel · D. M. Hayden · D. Coleman-Derr · T. A. Hill · O. D. Anderson · Y. Q. Gu (✉)
Genomics and Gene Discovery Research Unit, USDA-ARS
Western Regional Research Center,
800 Buchanan Street,
Albany, CA 94710, USA
email: ygu@pw.usda.gov

F. M. You · Y. Ma · J. Dvorak · M.-C. Luo (✉)
Department of Plant Sciences, University of California,
Davis, CA 95616, USA
e-mail: mcluo@ucdavis.edu

Introduction

The temperate grass *Brachypodium distachyon* (*Brachypodium*) is being developed into a model system for grasses (Draper et al. 2001), which currently lacks an appropriate model system. Rice is not an ideal model for the grasses because of its large size, demanding growth requirements, long generation time, and its adaptation to semi-aquatic tropical conditions. *Brachypodium* has a number of attributes similar to those that make *Arabidopsis* a good model system for dicotyledonous plants. *Brachypodium* is easy to grow, has a short generation time, is self-fertile, and has one of the smallest genomes of any grass species (~350 Mb). Recognition of the potential of *Brachypodium* as a model grass has recently been acknowledged by the US Department of Energy in their recommendation of *Brachypodium* as the basic model species for development of grasses as a feedstock for biofuels (http://genomicsgtl.energy.gov/biofuels/). To fully establish *Brachypodium* as a model system, a number of necessary methods and resources are being developed. Transformation by micro-

projectile bombardment and *Agrobacterium tumefaciens* have been reported (Christiansen et al. 2005; Vogel et al. 2006a). Large-insert genomic libraries using bacterial artificial chromosome (BAC) vectors have been constructed for *Brachypodium* (Hasterok et al. 2006; Huo et al. 2006) and the related perennial species *Brachypodium sylvaticum* (Foote et al. 2004). As yet, there has been a lack of sufficient DNA sequence information to infer the details about *Brachypodium* genome structure. Before this study, the available sequence resources were 20,449 expressed sequence tags (ESTs; Vogel et al. 2006b), 2,185 randomly sampling sequences of BAC ends (Huo et al. 2006), and 5 BAC clones from *B. sylvaticum* (Bossolini et al. 2007; Griffiths et al. 2006).

The genus *Brachypodium* and wheat both belong to the same subfamily (*Pooideae*), which diverged from the subfamily *Ehrhartoideae*, to which rice belongs, nearly 50 million years ago (Gaut 2002; Kellogg 2001). Therefore, in addition to serving as a general grass model, *Brachypodium* may serve particularly well for the Triticeae grasses, which include several important crops (wheat, barley, rye, and triticale) and other numerous important forage grasses. Analysis of *Brachypodium* ESTs identified many close matches to sequences derived from wheat and barley and placed *Brachypodium* near the base of the Triticeae branch of grasses (Vogel et al. 2006a; Vogel et al. 2006b). In addition, comparisons of orthologous segments of wheat, rice, and *B. sylvaticum* suggested that synteny between *B. sylvaticum* and the Triticeae was better than between either of them and rice (Bossolini et al. 2007; Griffiths et al. 2006).

The use of BAC end sequences (BES) in whole genome characterization was first proposed as a strategy for identifying overlapping clones during whole genome sequencing (Venter et al. 1996). BES also provide a sampling of the genome that can be used to estimate gene spacing, genome size as compared to known genomes (Hong et al. 2006), and the distribution of repetitive elements including transposable elements (TE), simple sequence repeats (SSR), and other classes of repeats (Hong et al. 2006; Lai et al. 2006; Mao et al. 2000; Paux et al. 2006). BES have also found use in sequence assemblies of plant genomes (Goff et al. 2002; The Rice Chromosome 10 Sequencing Consortium 2003). BES analyses have been carried out in a number of plants in the initial stages of genome characterization, including maize (Messing et al. 2004), Chinese cabbage (Hong et al. 2006), papaya (Lai et al. 2006), and wheat (Paux et al. 2006). BES are also rich sources of molecular markers (Paux et al. 2006; Tomkins et al. 2004) and have been used to infer the phylogenetic relationship between genomes. An example of the latter is the report by Lai et al. (2006) who found more synteny between papaya and poplar than either of them with *Arabidopsis*, although the genomes of papaya and *Arabidopsis* are more closely related.

To gain a sense of the genome structure of *Brachypodium*, we generated 64,696 BES – representing approximately 10.9% of the *Brachypodium* nuclear genome. Analysis of these BES provides a snapshot of the composition and organization of the *Brachypodium* genome, which allows an assessment of its phylogenetic relationships with several important cereal crops. Additionally, BES are invaluable molecular markers for physical mapping and sequencing of the *Brachypodium* genome.

Materials and methods

Brachypodium distachyon BAC libraries

Previously, we constructed two BAC libraries for *Brachypodium* using two restriction enzymes (*Hind*III and *Bam*HI; Huo et al. 2006). To increase the genome coverage for our ongoing *Brachypodium* physical mapping project, a second *Bam*HI BAC library was generated using the same strategy (Huo et al. 2006). A total of 36,864 BAC colonies of this library were picked into 96 384-well plates. The average insert size of the second *Bam*HI library was estimated to be 105 kb (data not shown), representing 9.9-fold haploid genome equivalents. For convenience, the two *Bam*HI libraries were combined. The combined *Bam*HI library contained 36,864 BAC clones (384-well plate numbers 1 to 96) from the first *Bam*HI BAC library and 36,864 BAC clones (384-well plate numbers 97 to 192) from the second *Bam*HI BAC library. The total genome coverage of the combined *Bam*HI library plus the *Hind*III library was approximately 29.2-fold haploid genome equivalents.

Sequencing of BAC ends

BAC clones of the *Hin*dIII and combined *Bam*HI BAC libraries were inoculated into 96-deep well blocks containing 1.2 ml/well 2xYT medium (Teknova, Hollister, CA) and grown overnight at 37°C with shaking at 300 rpm. The cells were harvested by centrifugation, and BAC DNA was purified using a REAL Prep 96 Plasmid Kit (Qiagen, Valencia, CA). For BAC-end sequencing, 5 μl of purified BAC DNA (~ 0.2 to 0.5 μg) was used in a sequencing reaction with BigDye v 3.1 (Applied Biosystems, Foster City, CA). Template DNA was sequenced from both directions with pCC1BAC/pIndigoBAC-5 Forward and Reverse End-Sequencing Primers (Epicentre,

Madison, WI). Electrophoresis of the sequencing reaction was carried out with a 3730xl DNA Analyzer (Applied Biosystems).

Sequence data processing

Sequence and quality files from trace files were read by the Phred program for base calling and quality trimming using a quality score of 20 (Ewing and Green 1998; Ewing et al. 1998). Vector sequences were masked using CROSS_ MATCH (http://www.genome.washington.edu), and the masked terminal vector sequences were removed. BES less than 100 bp were also removed from further analysis. The high-quality sequence data were then filtered for sequences contaminated with *Escherichia coli* or with plant organelle genomes based on matches to the wheat mitochondria (AP008982) and chloroplast (AB042240) sequences.

Sequence analysis

Processed BES were compared with several repeat databases including Triticeae repetitive (TREP) sequence database (http://wheat.pw.usda.gov/ITMI/Repeats/) (April, 2007), GIRI repeat database (http://www.girinst.org/) (April, 2007), and the TIGR plant repeat database (ftp://ftp.tigr.org/pub/data/TIGR_Plant_Repeats/) (April, 2007) using BLASTN and BLASTX at an *E* value cutoff of 10^{-5}. A survey of the composition and contents of *Brachypodium* repeat element sequences in BES was conducted using the RepeatMasker program (http://www.repeatmasker.org/). The BES were annotated based on their best match to the repeat database and categorized according to a reference repetitive element database (Repbase), which highlighted those annotated for the plant family Poaceae (Jurka et al. 2005).

For computational detection of SSR, BES were screened with all combination of di-, tri- and tetra-nucleotides using the SSR Search program (ftp://gramene.org/pub/gramene/software/scripts/ssr.pl). SSRs with dinucleotide, trinucleotide, and tetranucleotide motifs that span 12 or more nucleotides were recorded. As a comparison, the BES of wheat 3B (DX363346-DX382744), maize (http://www.genome.arizona.edu/stc/maize), rice (http://www.genome.clemson.edu/projects/rice), *Arabidopsis* (ftp://ftp.tigr.org/pub/data/a_thaliana/bac_end_sequences), soybean (CG812653-CG826126), and papaya (DX458351-DX502755) were directly downloaded and screened for SSRs using the same parameters with the same ssr.pl program.

For annotation of BES, repeat masked BES were also compared using BLASTN or BLASTX with non-redundant and dbEST database of NCBI (February, 2007), as well as UniProt (Ver. 9.7) of European Bioinformatics Institute database (February, 2007) to identify sequences similar to known genes. Gene ontology (GO) terms (April, 2007) were assigned to the BES that had significant matches to the UniProt reference sequences. Categories were assigned on the basis of biological, functional, and molecular annotations available from GO (http://www.genontology.org/).

Identification of unique *Brachypodium* repeat sequences

Self-BLASTN was performed on repeat-masked BES to identify sequences that had multiple strong matches to other BES with an *e*-value $<10^{-50}$. Sequences with a minimum of five matches were extracted and aligned by ClustalW analysis (Thompson et al. 1994). The consensus sequences with a minimal length of over 100 bp were retrieved from each alignment. The consensus sequences were compared with each other and aligned using CAP3 software (Huang and Madan 1999) to merge overlapping regions and to extend the sequences. These putative *Brachypodium* repetitive DNAs were BLASTN compared against known repetitive DNA databases to verify the uniqueness of the sequences. BLAST searches against nonredundant nucleotide, EST, and protein databases were also performed. A programming pipeline program was developed to implement the above procedure.

Phylogenetic analysis

To evaluate the evolutionary relationship of *Brachypodium* with six other grass species, rice (*Oryza sativa*), wheat (*Triticum aestivum*), barley (*Hordeum vulgare*), maize (*Zea mays*), sorghum (*Sorghum bicolor*), and sugarcane (*Saccharum officinarum*), the top matches of *Brachypodium* BES to ESTs of individual species were extracted. A total of 335 sequences that had a significant match in all the selected grass species were used in phylogenetic analysis. Multiple alignments using ClustalW (Thompson et al. 1994) were performed among matched sequences to reveal nucleotide polymorphisms. From completely aligned sequences, individual nucleotide sites displaying polymorphisms were collected to create a single combined dataset for phylogenetic analysis. This dataset included a total of 19,566 concatenated polymorphic sites. Five algorithms (DNAml, DNAmlk, DNAcomp, DNApars, and DNApenny) of phylogenetic analysis provided in Phylip package (version 3.6) were used to construct species trees (http://evolution.genetics.washington.edu/phylip.html). The resampling technique bootstrap was used to create 1,000 sampling sequence datasets using the bootseq program in the Phylip package,

and then consensus trees were obtained using the consensus program in the Phylip package.

Results

BAC-end sequencing and data processing

The two *Brachypodium* BAC libraries previously constructed using two different restriction enzymes, *Hin*dIII and *Bam*HI, represented 9.9-fold and 9.4-fold genome coverage, respectively (Huo et al. 2006). When combined with the second *Bam*HI library (9.9-fold) constructed in this study, the total genome coverage by the three libraries was 29.2-fold haploid genome equivalents. After preliminary screening for length and quality assessments (Lazo et al. 2004; Lazo et al. 2001), a total of 67,092 high-quality BES were generated. Among them, 2,396 (4%) sequences were eliminated after masking vector sequences and sequences contaminated with bacterial and organelle genomes using CROSS_ MATCH and BLASTN searches (Lazo et al. 2001). In total, 64,694 BES with a minimal size of 100 bases per BES and PHRED score ≥20 were generated for the *Brachypodium* nuclear genomes. The average size of the BES was 583 bp. The BES were deposited in the genome survey sequence section of the GenBank database with accessions from EI108890 to EI173585. Among the deposited BES, 35,082 were from the combined *Bam*HI library and 29,614 from the *Hin*dIII library. The total BES length was 38,240,859 bp with a GC content of 45.9%. Based on the genome size of 350 Mb estimated for *Brachypodium* (Bennett and Leitch 2005; Vogel et al. 2006a), the total BES was equivalent to 10.9% of the *Brachypodium* genome.

Analysis of repetitive DNA in BES

The small size of the *Brachypodium* genome implies a repetitive DNA content more similar to that of rice than that of the large genomes of maize, barley, or wheat. The 64,694 *Brachypodium* BES were first compared with several repetitive DNA databases to identify sequences with homology to characterized repetitive DNA classes (Table 1). Similar to other eukaryotic genomes, TE constitute a significant portion of the *Brachypodium* genome. Overall, 9,114 BES (14.1% of the total number of BES) had sequence homology to the Class I or RNA TEs (Table 1). The Class I TEs could be further classified into Ty1/copia (3426) and Ty3/gypsy LTR retrotransposon (4256), and LINE (1273) and SINE (159) non-LTR retrotransposons. Clearly, LTR retrotransposons outweighed non-LTR retrotransposons with respect to both the number of matches in BES and the percentage of the *Brachypodium* genome they comprised. The accumulated sequence length of LTR retrotransposons accounted for a total of 6.90% of the *Brachypodium* genome (350 Mb), while non-LTR retroelements only accounts for ~ 0.97% of the genome. Taken together, Class I TEs represented 7.87% of the genome.

Table 1 Occurrence and distribution of known repetitive DNA in the *Brachypodium* BAC end sequences

Class, subclass, group	No. of matches	No. of bases (bp)	Percentage of the genome (%)
Class I retrotransposon	9,114	3,003,303	7.87
LTR retroelement	7,682	2,633,957	6.90
Ty1/Copia	3,426	1,236,342	3.24
Ty3/Gypsy	4,256	1,397,615	3.66
Non-LTR retroelement	1,432	369,346	0.97
SINEs	159	24,462	0.06
LINE3	1,273	344,884	0.91
Class II DNA transposon	2,868	488,506	1.28
hobo-AC-Tam3	207	41,335	0.11
TC1-IS630-Pogo	1,693	192,262	0.50
En-Spm	518	165,392	0.44
MuD-IS905	317	69,584	0.18
Tourist/Harbinger	133	19,933	0.05
Unclassified	57	9,128	0.02
Total transposon elements	121,039	3,500,937	9.17
Small RNA	809	466,546	1.22
Simple repeats	1,899	110,754	0.29
Low complexity	2,706	116,855	0.31
Total known repetitive DNA	17,533	4,201,350	10.99

The next most abundant repeat DNA in the BES was Class II or DNA-mediated TEs. There were 2,868 BES (4.4% of the total number of BES) with homology to this type of repeat DNA elements, constituting approximately 1.28% of the *Brachypodium* genome (Table 1). The Class II TEs accounted for a sixth of the genome fraction represented by Class I TEs. In addition, 57 BES were identified that belong to unclassified plant repeat elements in the GIRI database. The total number of BES containing TE elements was 12,029 or 18.6% of the total BES. Other repeat sequences identified were ribosomal RNA genes (809), simple repeats (1,899), and low complexity DNA (2,706). Their accumulated sequence accounted for 1.82% of the genome. Taken together, our analyses of BES against existing repeat databases indicate that 11.0% of the genome corresponds to known repeat sequences (Table 1).

Unique *Brachypodium* repeat element sequences

A comparison with the existing repeat DNA databases allowed the masking of BES that have matches to the known repeat DNA sequences. However, repetitive DNA elements are known to evolve rapidly, and species- or genome-specific elements are likely present in the *Brachypodium* genome. To examine repeat DNA elements that are unique to the *Brachypodium* genome, the known repeat elements in BES were masked and then compared to other BES. Sequences that have multiple matches (more than five times) at high stringency (aligning at least 60 contiguous bases) were identified as putative repeat sequences. Identification of novel repetitive elements using computational approaches has been previously reported (Bao and Eddy 2002; Wiedmann et al. 2006). We employed a similar strategy to search for novel *Brachypodium* repetitive elements using the BES dataset. On the basis of sequence alignments of overlapping BES, we identified 294 consensus sequences with a size range from 148 to 3071 bp (Supplement 1). Because BES averaged 583 bp, the putative *Brachypodium* repetitive elements could be considerably longer. Search of other databases with the putative *Brachypodium* repetitive elements resulted in discovery of 19 protein-encoding genes undoubtedly representing large gene families (Table 2 and see Supplement 1). They were removed from the list of unique *Brachypodium* repeat element sequences. Nine sequences matched to the protein sequences of transposable elements, although no significant matches at the nucleotide level were detected. These sequences could represent repetitive elements whose coding sequences have diverged considerably in *Brachypodium*. The remaining 266 of the 294 putative repeat sequences had no significant matches in both BLASTN and BLASTX searches; they were termed unique *Brachypodium* repeat element sequences (*UBRES*).

The *UBRES* were used to analyze their prevalence in the *Brachypodium* genome by re-BLAST against the BES database. Because of the extended length of *UBRES* as compared to that of BES, more matches to *UBRES* were often found than in the initial search using a single BES containing repetitive sequence. *UBRES-2* had the highest number of matches to the BES database (511 at $e<10^{-25}$). Considering the genome coverage of 10.9% by the BES, there could be 4,700 copies of *UBRES-2* in the entire *Brachypodium* genome. By this estimation, *UBRES-4*, *UBRES-5*, *UBRES-6, UBRES-9*, *UBRES-10*, and *UBRES-62* each has over 1,000 copies in the genome. Over 100 *UBRES* have more than 100 copies in the genome. A total of 5,384 matches of *UBRES* were detected in the BES, accounting for 7.4% of the genome sequence based on *UBRES*-masked bases in the total bases of the BES. Adding these to the known repetitive DNA (11.0%; Table 1) showed that repetitive DNA accounts for ~18.4% in the *Brachypodium* genome.

Table 2 Analysis of unique *Brachypodium* repetitive element sequences

Annotation	No. of sequences	No. of matches in BES	Percent of the genome
Multi-gene families	19	214	0.22
Transposable proteins	9	144	0.34
No hit	266	5,384	7.10

Bare-1-related LTR retrotransposons in *Brachypodium*

Long terminal repeat (LTR) retrotransposable elements are known to play a significant role in the evolution dynamics of plant genomes (Casacuberta and Santiago 2003; Vitte and Bennetzen 2006; Vitte and Panaud 2005). The LTRs are direct sequence repeats that flank the internal region, which encodes genes of structural and enzymatic proteins required for replication and transposition. The unique LTR region, ranging from less than 100 bp to several kb in size, can account for a considerable portion of LTR retroelements (Wicker and Keller 2007). LTR sequences are usually not conserved among distantly related species. They could be highly divergent even among the same type of retrotransposons, such as Ty3/gypsy retrotransposons, within the same plant species, although the internal coding sequences might show considerably similarity (85–90% identity). Therefore, LTR retrotransposons can be subdivided based on the sequence identity in the LTR region (Wicker and Keller 2007). Search of homology in *Brachypodium* BES with Triticeae sequences of characterized LTR retroelements in the TREP database (http://wheat.pw.usda.gov/ggpages/ITMI/Repeats/) yielded matches to the LTR regions of three closely related retroelements, *Bare-1*, *WIS*, and *Angela*. *Bare-1* was first identified in barley (Manninen and Schulman 1993), and its LTR sequences share a high nucleotide identity to those of wheat *WIS* and *Angela* retrotransposons (Muniz et al. 2001; SanMiguel et al. 2002). Retrieved *Brachypodium* sequences using the LTR of these three retrotransposon were 99% identical. We named the consensus sequence of 1,713 bp *BRES-1* LTR (*Brachypodium* repetitive element sequence-1).

The complete *BRES-1* sequence shares ~72% nucleotide identity with the LTR sequences of barley *Bare-1* and wheat *WIS* and *Angela* retroelements. The perfect inverted repeat (TGTTGG-CCAACA) at the termini of LTRs between *Bare-1* and *BRES-1* elements remained conserved, although dot matrix analysis of the two LTRs revealed sequence divergence in some other regions (Fig. 1a). For

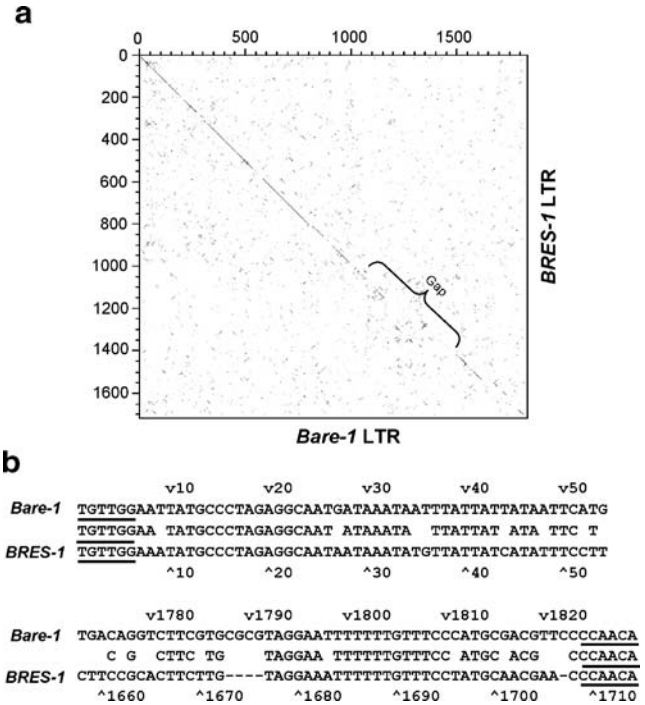

Fig. 1 **a** *Bare*-1-like LTR retrotransposon in *Brachypodium*. A dot matrix analysis on the full-length LTR sequences of the *Bare-1* retrotransposon from barley and a *Bare-1*-like retrotransposon from *Brachypodium* (*BRES-1*) were performed using window size of 40 bp and 60% of minimal match. The *diagonal lines* represent regions where the sequences shared similarity between the two LTRs. The gap regions represent divergent sequences. A gap region with a size of ~500 bp is indicated. **b** Sequence alignment of wheat and *Brachypodium Bare-1* elements. The first 54 and last 58 bases of the LTR regions from barley *Bare-1* and *Brachypodium BRES-1* elements were aligned with ClustalW program. A consensus sequence is shown between the two LTRs. The perfect repeats at the terminal ends of the LTR are *underlined*

example, a 500-bp segment at the position of 1,000–1,500 bp in the LTR region is not conserved between *Bare-1* and *BRES-1*(Fig. 1a. The 3′ end of the LTRs are more divergent than 5′ end, although the last 50 bps are quite conserved (Fig. 1b). In addition, the *BRES-1* LTR was about 100-bp shorter than the intact *Bare-1* LTR. A BLAST search using this *BRES-1* LTR sequence retrieved 250 *Brachypodium* BES with an $e<10^{-50}$. By estimation, there are approximately 2,293 copies of this LTR sequence in the *Brachypodium* genome. As each intact LTR retrotransposon contains two LTRs, there are about 1,146 *BRES-1* retroelements in the genome. Assuming that the size of the internal coding region of *BRES-1* is comparable to that of *Bare-1* (~8.4 kb), the *BRES-1* elements with an estimated size of 11,858 bps would contribute around 13.6 Mb sequence to the *Brachypodium* genome. However, this is likely to be an overestimate, as many LTR retrotransposons exit as solo LTRs due to the high frequencies of homologous recombination between two LTRs, and illegitimate recombination often resulted in fragmented retrotransposon elements with smaller sizes (Devos et al. 2002; Ma et al. 2004).

In addition, when the LTR sequences of *Ty3*/gypsy retrotransposons from the TREP database were used in a BLAST search against the BES, no significant matches were found in the *Brachypodium* genome sequence. However, our analysis of known repetitive elements in the BES showed that Ty3/gypsy retroelements constitute about 3.7% of the *Brachypodium* genome (Table 1). We found that the matches to Ty3/gypsy retroelements often come from internal domains containing conserved genes required for transposon amplification and that the LTR sequences of *Brachypodium Ty3*/gypsy elements are considerably divergent from those of Triticeae *Ty3*/gypsy elements.

Analysis of SSR

A total of 10,144 SSRs were identified in the 38.2 Mb of the *Brachypodium* BES. Of these SSRs, 1,291 (12.7%), 4,527 (44.6%), and 4,326 (42.6%) represent dinucleotide, trinucleotide, and tetranucleotide motifs, respectively (Table 3). The most abundant repeat motifs were trinuclotide and tetranuclotide repeats. The frequency of SSRs derived from *Brachypodium* BAC-end sequence was about one SSR per 3.9 kb of genomic sequence. To compare the frequency and distribution of BES-derived SSRs in different species, BES totaling 10.8 Mb from wheat 3B, 84.6 Mb from rice, 38.2 Mb from maize, 27.5 Mb from *Arabidopsis*, 9.9 Mb from soybean, and 18.7 Mb from papaya were downloaded for screening for SSRs using the same computational algorithm and parameters, which allowed us to make a fair

Table 3 The SSR frequency and distribution in BES from different plant species

	Brachypodium (350 Mb)	Wheat (3B)[a] (1,000 Mb)	Rice (389 Mb)	Maize (2,365 Mb)	*Arabidopsis* (125 Mb)	Soybean (1,115 Mb)	Papaya (372 Mb)
Dinucleotides	1,291 (12.7%)	324 (18.2%)	4,554 (23.4%)	814 (17.7%)	1,538 (22.9%)	794 (29.6%)	3,841 (46.7%)
Trinucleotides	4,527 (44.6%)	943 (52.9%)	7,538 (38.7%)	2,297 (49.9%)	2,921 (43.5%)	854 (31.8%)	1,968 (23.9%)
Tetranucleotides	4,326 (42.6%)	515 (28.9%)	7,388 (37.9%)	1,496 (32.5%)	2,249 (33.5%)	1,038 (38.6%)	2,415 (29.4%)
SSR frequency[b]	3.9 kb	6.1 kb	4.3 kb	8.3 kb	4.1 kb	3.7 kb	2.2 kb

[a] Wheat 3B chromosome
[b] Average estimated distance between SSRs

comparison of different SSR motifs among different plant genomes. In these seven species, small genomes tended to have higher frequencies of SSRs as compared to large genomes, such as wheat and maize. This is in agreement with the previous finding that there is significant association of SSRs with low-copy fraction of DNA in plant genomes and the SSR frequency is inversely related to the proportion of repetitive DNA, particularly to LTR retrotransposons (Morgante et al. 2002). However, papaya had the highest SSR frequency despite the fact that its genome size is more than twice that of *Arabidopsis*.

We observed significant variations in the frequency of SSR motifs in different genomes. In papaya and soybean, dinucleotide or tetranucleotide repeats were the most abundant motifs, while in the other species, trinucleotide repeats were the most frequent motifs (Table 3). When different dinucleotide repeats in the same species were compared, the AG/CT motifs were the most abundant repeats in *Brachypoidum* and *Arabidopsis*, representing 55.2 and 52.5% of all SSR, respectively, while the AT/TA repeats were the most abundant in papaya and soybean (Fig. 2). For the trinucleotides repeats, the GC-rich repeats accounted for more than 60% of all trinucleotide repeats in *Brachypodium* and rice (67.4 and 65.0%), and AT-rich trinucleotide repeats were the majority in the other five plant species. Previous studies have shown that the SSR frequencies are higher in transcribed regions than in genomic DNA, and in the transcribed regions, GC-rich trinucleotide repeats are the majority in monocotyledonous plants (Morgante et al. 2002). The high frequency of the GC-rich trinucleotide repeats in *Brachypodium* and rice may be due to the fact that a relatively high percentage of their compact genomes are transcribed. This also provides a plausible explanation of the observation that the GC-rich trinucleotide repeats are much less frequent than the AT-rich repeats in the large and highly repetitive genomes of maize and wheat. Therefore, it is possible that AT-rich repeats are higher in repetitive DNA from wheat and maize. However, in dicotyledonous, AT-rich repeats are the predominant motif in the trinucleotide repeats (Fig. 2). It appears that GC-rich trinucleotide repeats are not associated with the coding sequences even in dicotyledonous with small genomes.

The ubiquity and variability of SSRs make these sequences particularly useful for creating PCR-based markers for genetic mapping. Thus, the 10,144 SSRs we identified will be a valuable mapping resource. We have already used 480 of these SSRs to identify useful polymorphisms between six inbred *Brachypodium* lines. About 46 out of 480 SSRs were found to be consistent and robust enough to use for genetic mapping of a *Brachypodium* mapping population (unpublished data).

Sequence comparison with different grass EST databases

After masking the repeat DNA sequences, the remaining BES (55,221) were compared to the EST database to estimate the transcribed portion of the *Brachypodium* genome. Transposon-related genes are known to be present in the EST database and can skew the results if the BES contain high percentage of TE sequences owing to the conservation of transcribed TEs between species (Messing et al. 2004). As repeat masked BES were used in our analysis, any matches to ESTs are likely to represent transcribed portions of gene sequences in the genome and not transposon-related genes. Using a cutoff of $e<10^{-5}$, 37.5% of the BES (20,707) had matches against the dbEST database. Even at $e<10^{-25}$, 25.3% of BES (13,970) found matches, representing 8.1 Mb of the BES based on the average length of 583 bp per BES. Based on the stringency of $e<10^{-25}$, the estimated transcribed portion of the *Brachypodium* genome is 74.2 Mb, given 10.9% of the genome coverage by BES. In other words, approximately, 21.2% of the genome may represent coding regions.

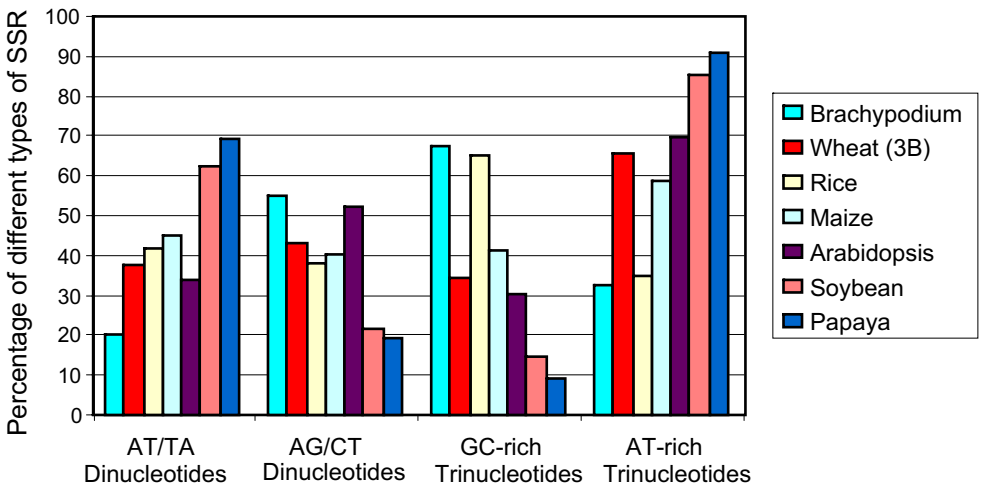

Fig. 2 Frequencies of dinucleotide and trinucleotide repeat motifs in the BES from seven plant species. The BES from different species were downloaded from the NCBI. Different types of dinucleotide and trinucleotide repeats were searched for each grass species. The frequencies of different types of motifs in dinucleotide or trinucleotide repeats were calculated for each species

Assuming that the average size of *Brachypodium* genes is similar to that of rice (~3.0 kb), the estimated number of genes is ~25,000 or one gene per 14.0 kb of the *Brachypodium* genomic sequence. The overall gene density estimated here is lower than the gene density (approximately one gene per 8.0 kb) reported for the *B. sylvaticum* genome based on the analysis of a sequenced 371-kb region (Bossolini et al. 2007). This may reflect the considerable variations of gene density at different genomic regions.

The repeat masked BES were also BLASTN compared against individual EST data sets from different grass species. Such a comparison allowed us to determine the degree of sequence similarity to different grass species. Seven grass species including *Brachypodium* and six other grass species (rice, maize, wheat, barley, sorghum, and sugarcane) were included in the analysis (Fig. 3). In this study, the BES data was BLASTN compared against the NCBI dbEST database, as EST collections were available for each species, whereas only rice had a complete genome sequence. The number of matches of BES against individual EST databases at different *E* values was determined and plotted to assist estimating the relatedness of *Brachypodium* with other grass species. If a species is closely related to *Brachypodium*, a greater number of matches at more stringent *E* values would be expected, assuming that the EST collections are large enough and well represented in terms of tissue specificity and developmental stages. When comparisons were performed, the sizes of EST collections for wheat (1,050,131), rice (1,211,154), and maize (1,161,241) were comparable but larger than those of barley (437,738), sorghum (204,308), and sugarcane (246,301). *Brachypodium* had the smallest EST collection (20,449), but it had the highest number of matches with BES at a high stringency ($e<10^{-70}$). This can be explained by the fact that a matching score will be highest when the BES found its corresponding *Brachypodium* EST. However, matches to *Brachypodium* EST at a low stringency were present (Fig. 3), suggesting that only a related sequence, i.e., a diverged paralog, is identified. The corresponding gene sequence is not present due largely to the small EST collection for *Brachypodium*. Therefore, when the *E* value cutoff is increased to $e<10^{-60}$, the matches to wheat ESTs outnumber those to *Brachypodium* ESTs (Fig. 3). This reflects the significant difference in size of the wheat and *Brachypodium* EST collection. In a simulated analysis in which only a small set of wheat ESTs, similar to the size of *Brachypodium* EST collection, was randomly selected for BLASTN comparison, the number of matches of BES to wheat was fewer than those to *Brachypodium* at any *E* value cutoffs (data not shown). Nevertheless, wheat had a higher number of matches at any *E* value cutoff than rice or maize, despite the similar sizes of their EST collection (Fig. 3). Barley, a closely related species to wheat, also showed a higher number of matches than rice or maize at highly stringent *E* value cutoffs, although its EST collection is less than half the size of these species. These results further support a closer relationship between *Brachypodium* and the Triticeae species than the remaining grasses studied.

GO annotations

The repeat masked BES were also screened against the UniProt database for functional categorization of the predicted protein sequences derived from the BES. A total

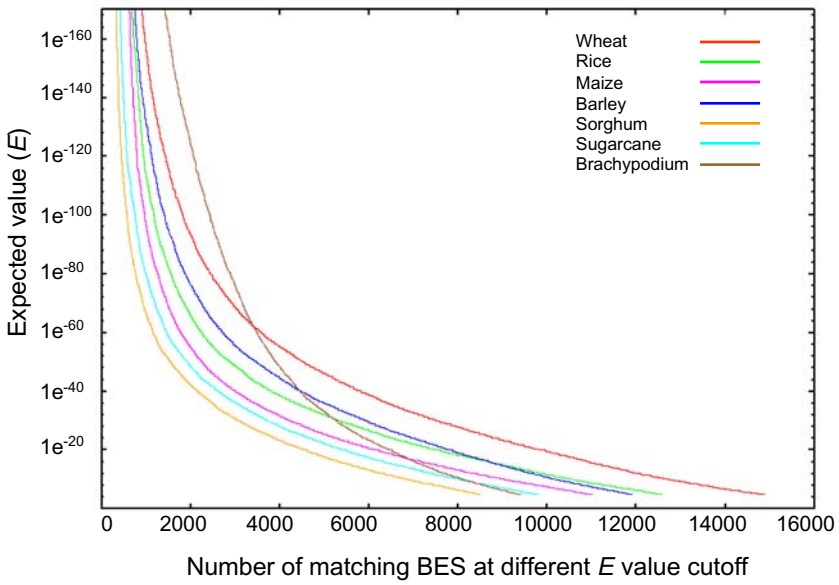

Fig. 3 BLAST search of *Brachypodium* BES against EST collections from different grass species. Repeat masked BES were compared against individual EST collections from grass species using BLASTN. The numbers of BES that match EST collections at *E* cutoff equal to or less than a specific value were calculated and plotted for each species using indicated *color lines*

of 12,113 (24%) of the BES had BLASTX matches at e^{-10} or better. A GO translation table provided candidate GO annotations. GO terms were rooted using accessions for biological (GO:0008150), cellular (GO:0005575), and molecular (GO:0003674) categories. Matches of BES to GO terms were assigned and fell into unique biological (1,886), cellular (501), and molecular (1,444) categories. Distribution of molecular GO terms is presented in Table 4. The top four GO categorizations were associated with transferase activity and binding to protein, ions, or nucleic acids. The abundance for each functional category for *Brachypodium* was comparable to findings in other plant species using BES data (Messing et al., 2004; Hong et al., 2006). Further description of the terms may be obtained from the *Brachypodium* data resource available at http://brachypodium.pw.usda.gov.

Phylogenetic analysis

Phylogenetic analyses based on DNA sequences from a single locus or a few loci can yield contradictory results, primarily due to limited character sampling (Rokas et al. 2003). It has been proposed that a data set of sequence information derived from a large number of genes can greatly improve phylogenetic accuracy (Rokas and Carroll 2005). To better evaluate the evolutionary relationship of *Brachypodium* with other grass species, we identified 335 BES sequences that have significant matches ($e<10^{-50}$) to ESTs from seven grass species (see Supplement 2). In cases where the BES had multiple matches in the EST database of individual species, only the top match was selected for further analysis. ClustalW was used to align sequences for each individual gene sequence. Aligned sites that displayed nucleotide polymorphisms were extracted and concatenated for each species. The concatenated sequences contain a total of 19,566 variable sites. Phylogenetic analysis of the concatenated sequences with five algorithms (DNAml, DNAmlk, DNAcomp, DNApars, and DNApenny) created a consistent topology from 1,000 bootstrapped datasets (Fig. 4). *Brachypodium* was shown to be much closer evolutionarily to wheat and barley than the other grass species.

Table 4 Distribution of functional classes of genes tagged to *Brachypodium* BES

No.	Category	Percent of protein-coding BES
1	Transferase activity	12.6
2	Protein binding	12.0
3	Ion binding	11.3
4	Nucleic acid binding	10.4
5	Hydrolase activity	8.2
6	Oxidoreductase activity	6.8
7	Signal transducer activity	4.0
8	Lipid binding	1.8
9	Carbohydrate binding	1.7
10	Ion transporter activity	1.7
11	Lyase activity	1.6
12	Steroid binding	1.4
13	Carrier binding	1.4
14	Transcription factor activity	1.3
15	Cofactor binding	1.2
16	Helicase activity	0.8
17	ATPase activity	0.8
18	Tetrapryrole binding	0.7
19	Chromatin binding	0.6
20	Isomerase activity	0.6
21	Structural constituent of ribosome	0.6
22	Translation factor activity	0.5
23	Small protein conjugating enzyme activity	0.4
24	Other GO term	17.6

Discussion

In the present study, about 10.9% of the *Brachypodium* genome sequence derived from random BAC ends was analyzed to provide a first view of its genome structure and organization. The small genome size of *Brachypodium* was shown to be accompanied by low content of repetitive DNA, particularly transposable elements, which is the major component of large plant genomes such as those of wheat and maize. Our phylogenetic study using a genome-wide approach provided further evidence that *Brachypodium* has a closer evolutionary relationship with the Triticeae as compared to rice, maize, sorghum, and sugarcane. The results presented in this paper support development of *Brachypodium* to serve as an alternative model species for important grass species with the completion of the full genome sequence.

Repetitive elements in *Brachypodium*

BES represent relatively random and unbiased samples of sequence composition of a genome. The present analyses show that the *Brachypodium* nuclear genome contains 18.3% of repetitive sequences, with 11.0% homology to known repetitive DNA elements and 7.3% unique *Brachypodium* repetitive element sequences. This is consistent with the previous estimation that the *Brachypodium* genome contains about 15% highly repetitive DNA (Catalan et al. 1995). The data indicate that repetitive DNA content in *Brachypodium* is intermediate between rice (35%; IRGSP 2005) and *Arabidopsis* (10%; The Arabidopsis Genome Initiative 2000). This result is expected considering that the content of repetitive elements is

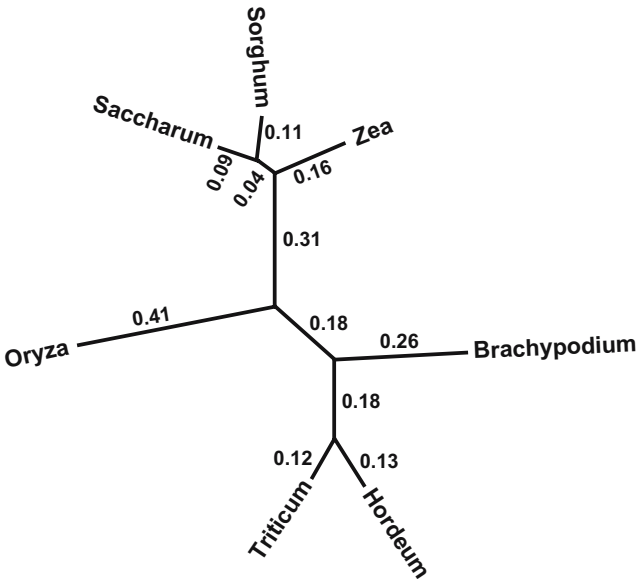

Fig. 4 Phylogenetic tree of *Brachypodium* and six other grasses. A consistent tree topology was created using five algorithms, DNAml, DNAmlk, DNApars, DNAcomp, and DNApenny, of phylogenetic analysis. The species tree shown here was created using DNAml (maximum likelihood) algorithm. The values on each branch represent the branch length estimated by maximum likelihood algorithm based on the raw dataset. For each of all algorithms, all trees based on 1,000 bootstrapped samples showed the same topology

directly correlated with genome size (Hawkins et al. 2006; Vitte and Bennetzen 2006), and the genome size of *Brachypodium* is likely between rice and *Arabidopsis* (Bennett and Leitch 2005; Vogel et al. 2006a). It is worth noting that the repetitive DNA content in *B. sylvaticum* is estimated to be 29% based on the sequence of a 372-kb region (Bossolini et al. 2007), which is considerably higher than that in the genome of *B. distachyon* (18.3%). However, the genome size of *B. sylvaticum* (470 Mb; Foote et al., 2004) is greater than that of *B. distachyon* (350 Mb), which might explain the difference of repetitive DNA content in the two genomes. The small size of the *B. distachyon* genome is one of the important attributes that makes it a model species for cereal crops with large and complex genomes, such as barely, maize, and wheat.

Repetitive elements comprise significant portions of most eukaryotic genomes and are increasingly known to impact genomic function (Akhunov et al. 2007; Casacuberta and Santiago 2003; Vitte and Bennetzen 2006; Wicker and Keller 2007). Understanding the ubiquity and importance of repetitive elements reinforces to the need to identify and annotate them in the genome. Analysis of TE distribution in the BES dataset revealed that Class I TEs significantly outweigh Class II TEs in the *Brachypodium* genome, as the number of Class I TEs is three times greater than that of Class II TEs and the genome fraction of Class I TEs (7.91%) is six times greater than that of Class I TEs (1.35%; Table 1). This contrasts from rice, for which the Class II TEs outnumber Class I TEs (IRGSP 2005) and the genome fractions occupied by the Classes I and II TEs are 19.35 and 12.96%, respectively. It seems that the different repetitive DNA contents of *Brachypodium* and rice can be largely explained by different contents of Class II elements. The genome fraction of Class II TEs in *Brachypodium* is also much lower than their fractions in the wheat D genome (11.6–13.0%; Li et al. 2004; Paux et al. 2006) but is more similar to that in the maize genome (0.9–1.3%; Messing et al. 2004; Paux et al. 2006). It appears that the proportion of genome represented by Class II TEs is variable among grass genomes and is not directly correlated to genome size. For example, the proportion of genome containing Class II TEs are comparable in the wheat D and rice genomes, although their sizes are dramatically different (389 Mb of the rice genome (IRGSP 2005) vs 4,000 Mb of the wheat D genome (Li et al. 2004)). Although the contribution of Class I elements, particularly LTR retrotransposons, to genome expansion has been well documented, the cause of the significant variations of Class II elements in different genomes remains unclear. It is possible that Class II TEs in different genomes may display differential activities in transposition or have different rates of turnover, resulting in distinct distributions during genome evolution. Therefore, it would be interesting to compare more closely aspects of rice and *Brachypodium* genomes, i.e., the genome sizes are comparable but the contents of Class II elements are significantly different.

Our approach to identify unique *Brachypodium* repetitive element sequence (*UBRES*) proved to be useful and efficient. The 265 *UBRES* identified in this study comprise approximately 7.4% of the *Brachypodium* genome. The extended length of *UBRES* from BES allows us to estimate the copy number for each *UBRES* in the genome. The results from screening *Brachypodium* BAC filters and genomic Southern hybridization using the *UBRES-4* probe provided biological evidence to support the in silico analysis of repetitive elements in our study (Supplement 3). In addition, several of *UBRES* were also found to be present in the orthologous *Lr34* and *Ph1* regions of the *B. sylvaticum* genome (Bossolini et al., 2007; Griffiths et al., 2006). It is worth noting that the *UBRES* identified might cover only portions of the intact elements, particularly for LTR retrotransposons. A LTR retrotransposon could have a unique LTR sequence, although its internal coding regions might be very similar to well-characterized transposon-related genes. In this case, the internal region may have been masked before the analysis of *UBRES*; only the LTR sequence will be retrieved. We found that *UBRES*-13 represents the LTR sequence of a *copia*-like retrotransposon as revealed by sequencing and annotation of random *Brachypodium* BAC clones (unpublished data). In addition, *UBRES* are useful in identification of repetitive elements that

are not present in the existing databases. When the *UBRES* were used to search BAC sequences derived from ten random BAC clones, BACs containing up to 34% of the *URBRES* were identified (unpublished data). With the genome sequence of *Brachypodium* available in the near future, identification and understanding of the evolution of *UBRES* will greatly improve the annotation of the genome.

The Triticeae Ty1/copia retrotransposons, *Bare-1*, *WIS*, and *Angela* are widely dispersed repetitive DNA elements in wheat and barley genomes (Muniz et al. 2001; SanMiguel et al. 2002; Suoniemi et al. 1996). Our results indicate that a closely related retroelement *(BRES-1)* is also abundant in the *Brachypodium* genome. Rice repeat databases were searched with the LTR sequence of *BRES-1*. Weak homology was detected with the rice *RIRE1* retrotransposon (~50% nucleotide identity; data not shown). The LTR sequence of the rice *RIRE1* retrotransposon has been shown to have poor homology with those of *Bare-1* and *WIS* elements, although their internal coding regions displayed good sequence conservation (Noma et al. 1997). The poor sequence homology in the LTR region can be explained by the high mutation rate in the LTR region due to lack of selection pressure (Petrov 2001). Taken together, it seems that Ty1/copia retrotransposons with LTR sequences related to those of *Bare-1* and *BRES-1* are ancient elements that existed before the divergence of Triticeae and rice. Therefore, the higher sequence conservation in the LTR region between *Brachypodium* and Triticeae was not caused by the horizontal sequence transfer as observed for some retrotransposons. It is more likely that *Brachypodium* and Triticeae are more closely related evolutionarily.

Phylogenetic relationship of *Brachypodium* with other grass genomes

Three lines of evidence support the conclusion that *Brachypodium* is closely related to wheat and barley. First, the *Bare-1*-like LTR retrotransposable element in the *Brachypodium* genome (*BRES-1*) has higher sequence identity in the LTR region with those of Triticeae *Bare-1*, *Wis-2*, and *Angela* elements than with the LTR sequence of a related rice retrotransposon, *RIRE1*. Second, the homology search between BES and individual EST collections from several grasses indicate that *Brachypodium* has a greater number of matches and higher match scores to wheat and barley than to rice or maize. Third, our phylogenetic analysis based on a large dataset of polymorphisms placed *Brachypodium* near the base of the Triticeae lineage.

A comprehensive study evolutionarily among *Brachypodium* and other grasses is important to maximize the utility of *Brachypodium* as a model system. Preceding molecular phylogenetic analyses have shown that the genus *Brachypodium* belongs to a sister group to the four major temperate grass tribes (*Triticeae*, *Aveneae*, *Poeae*, and *Bromeae*), which includes the majority of important temperate cereals and forage grasses (Kellogg 2001). However, one of most notable difficulties in molecular phylogenetic study is the widespread occurrence of incongruence between alternative phylogenies generated from single-gene data sets, as such strategies suffer from low resolution and marginal statistical support due to limited character sampling. It is known that individual gene genealogies could differ from each other and from the organismal phylogeny (gene-tree vs species-tree; Pamilo and Nei 1988). In fact, this has been observed in *Brachypodium* where phylogenetic analysis of individual genes resulted in different topologies (Vogel et al. 2006b). To overcome this limitation, it is necessary to construct a phylogeny using multiple genes as was done for *Brachypodium* using the partial sequence of 20 genes (Vogel et al. 2006b). In this study, we greatly increase the number of genes sampled by employing a phylogenomic approach to create a concatenated sequence dataset to determine the evolutionary relationship among seven grass species. The use of genome-wide data sets provide power not only in testing specific phylogenetic hypotheses but also in precise reconstruction of the historical association of all the taxa analyzed (Delsuc et al. 2005). Because a large number of sequences were used for the phylogenetic analysis, any bias contributed by a fraction of individual loci sampled will be minimized in the calculations. An advantage of this method is that it can directly use existing sequence collections. This method might not be applicable to species in which there are a low number of ESTs, as the chance of finding orthologous sequences will be greatly reduced and the likelihood of using the non-orthologous sequences in the analysis will be increased. For this reason, we only used seven grass species in our analysis; the inclusion of species with small EST dataset could greatly reduce sampling size due to the low probability of identifying the orthologous loci. Our analysis based on 335 gene sequences with a total of 19,566 concatenated polymorphic sites provides a robust phylogenetic tree among the species analyzed. This tree is also consistent with previous trees based on smaller datasets (Kellogg 2001; Vogel et al. 2006b).

Brachypodium genomics resource

Brachypodium has been proposed as a model system for temperate grasses and biofuel research (Bossolini et al. 2007; Draper et al. 2001; Garvin et al. 2007; Vogel et al. 2006b). With the *Brachypodium* whole genome sequencing and large scale EST production both underway (http://www.igi.doe.gov/sequencing/cspseqplans2007.html), *Brachypodium* is becoming an attractive system for a wide range of

biological research objectives. In just a few years, the diverse array of genomic resources and technologies that has been developed for *Brachypodium* will play a major role in optimizing the ability to use *Brachypodium* as a model system (Garvin et al. 2007). The large scale sequencing of BAC ends reported in this paper not only facilitates the first glimpse of this small genome structure but also contributes to the increasing list of useful genomic resources for this species. A website (http://brachypodium.pw.usda.gov) is available for local BLAST searches against the BES dataset, for *UBRES* generated from the BES, and for the SSR markers discovered in the BES, including flanking primers for genetic mapping in *Brachypodium*. Moreover, BAC clones with available BES will be useful for anchoring of the BAC clones and BAC contigs on a genetic map, facilitating the construction of a physical map of *Brachypodium* genome and rapid isolation of candidate genes of interest. BAC clones, BAC libraries used for generating BES, and membranes for screening the BAC libraries can be ordered from a website (http://wheat.pw.usda.gov/wgc/).

Acknowledgement We thank R. Naderi and C.X. Wang for help in BAC DNA preparation and C.C. Crossman for technical assistance in BAC-end sequencing. This work was support in part by the United State Department of Agriculture, Agriculture Research Service CRIS 532502100-000, 532502100-011, and 532521000-13.

References

Akhunov ED, Akhunova AR, Dvorak J (2007) Mechanisms and rates of birth and death of dispersed duplicated genes during the evolution of a multigene family in diploid and tetraploid wheats. Mol Biol Evol 24:539–550

Bao Z, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 12:1269–1276

Bennett MD, Leitch IJ (2005) Nuclear DNA amounts in angiosperms: progress, problems and prospects. Ann Bot 95:45–90

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

Casacuberta JM, Santiago N (2003) Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. Gene 311:1–11

Catalan P, Shi Y, Armstrong L, Draper J, Stace CA (1995) Molecular phylogeny of the grass genus *Brachypodium* P-Beauv based on RFLP and RAPD analysis. Bot J Linn Soc 117:263–280

Christiansen P, Andersen CH, Didion T, Folling M, Nielsen KK (2005) A rapid and efficient transformation protocol for the grass *Brachypodium distachyon*. Plant Cell Rep 23:751–758

Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6:361–375

Devos KM, Brown JK, Bennetzen JL (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res 12:1075–1079

Draper J, Mur LA, Jenkins G, Ghosh-Biswas GC, Bablak P, Hasterok R, Routledge AP (2001) *Brachypodium distachyon*. A new model system for functional genomics in grasses. Plant Physiol 127:1539–1555

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186–194

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175–185

Foote TN, Griffiths S, Allouis S, Moore G (2004) Construction and analysis of a BAC library in the grass *Brachypodium sylvaticum*: its use as a tool to bridge the gap between rice and wheat in elucidating gene content. Funct Integr Genomics 4:26–33

Garvin DF, Gu YQ, Hasterok R, Hazen SP, Jenkins G, Mockler TC, Mur AL, Vogel JP (2007) Development of genetic research resources for *Brachypodium distachyon*, a new model system for grass crop research. The Plant Genome (In press)

Gaut BS (2002) Evolutionary dynamics of grass genomes. New Phytologist 154:15–28

Goff SA, Ricke D, Lan T-H, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). Science 296:92–100

Griffiths S, Sharp R, Foote TN, Bertin I, Wanous M, Reader S, Colas I, Moore G (2006) Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat. Nature 439:749–752

Hasterok R, Marasek A, Donnison IS, Armstead I, Thomas A, King IP, Wolny E, Idziak D, Draper J, Jenkins G (2006) Alignment of the genomes of *Brachypodium distachyon* and temperate cereals and grasses using bacterial artificial chromosome landing with fluorescence in situ hybridization. Genetics 173:349–362

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Res 16:1252–1261

Hong CP, Plaha P, Koo DH, Yang TJ, Choi SR, Lee YK, Uhm T, Bang JW, Edwards D, Bancroft I, Park BS, Lee J, Lim YP (2006) A survey of the *Brassica rapa* genome by BAC-end sequence analysis and comparison with *Arabidopsis thaliana*. Mol Cells 22:300–307

Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9:868–877

Huo N, Gu YQ, Lazo GR, Vogel JP, Coleman-Derr D, Luo MC, Thilmony R, Garvin DF, Anderson OD (2006) Construction and characterization of two BAC libraries from *Brachypodium distachyon*, a new model for grass genomics. Genome 49:1099–1108

IRGSP (2005) The map-based sequence of the rice genome. Nature 436:793–800

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467

Kellogg EA (2001) Evolutionary history of the grasses. Plant Physiol 125:1198–1205

Lai CW, Yu Q, Hou S, Skelton RL, Jones MR, Lewis KL, Murray J, Eustice M, Guan P, Agbayani R, Moore PH, Ming R, Presting GG (2006) Analysis of papaya BAC end sequences reveals first insights into the organization of a fruit tree genome. Mol Genet Genomics 276:1–12

Lazo GR, Chao S, Hummel DD, Edwards H, Crossman CC, Lui N, Matthews DE, Carollo VL, Hane DL, You FM et al (2004) Development of an expressed sequence tag (EST) resource for wheat (*Triticum aestivum* L.): EST generation, unigene analysis, probe selection and bioinformatics for a 16,000-locus bin-delineated map. Genetics 168:585–593

Lazo GR, Tong J, Miller R, Hsia C, Rausch C, Kang Y, Anderson OD (2001) Software scripts for quality checking of high-throughput nucleic acid sequencers. Biotechniques 30:1300–1305

Li W, Zhang P, Fellers JP, Friebe B, Gill BS (2004) Sequence composition, organization, and evolution of the core Triticeae genome. Plant J 40:500–511

Ma J, Devos KM, Bennetzen JL (2004) Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome Res 14:860–869

Manninen I, Schulman AH (1993) BARE-1, a copia-like retroelement in barley (*Hordeum vulgare* L.). Plant Mol Biol 22:829–846

Mao L, Wood TC, Yu Y, Budiman MA, Tomkins J, Woo S-s, Sasinowski M, Presting G, Frisch D, Goff S, Dean RA, Wing RA (2000) Rice transposable elements: a survey of 73,000 sequence-tagged-connectors. Genome Res 10:982–990

Messing J, Bharti AK, Karlowski WM, Gundlach H, Kim HR, Yu Y, Wei F, Fuks G, Soderlund CA, Mayer KFX, Wing RA (2004) Sequence composition and genome organization of maize. Proc Natl Acad Sci USA 101:14349–14354

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30:194–200

Muniz LM, Cuadrado A, Jouve N, Gonzalez JM (2001) The detection, cloning, and characterisation of WIS 2-1A retrotransposon-like sequences in *Triticum aestivum* L. and Triticosecale Wittmack and an examination of their evolution in related Triticeae. Genome 44:979–989

Noma K, Nakajima R, Ohtsubo H, Ohtsubo E (1997) RIRE1, a retrotransposon from wild rice *Oryza australiensis*. Genes Genet Syst 72:131–140

Pamilo P, Nei M (1988) Relationships between gene trees and species trees. Mol Biol Evol 5:568–583

Paux E, Roger D, Badaeva E, Gay G, Bernard M, Sourdille P, Feuillet C (2006) Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. Plant J 48:463–474

Petrov DA (2001) Evolution of genome size: new approaches to an old problem. Trends Genet 17:23–28

Rokas A, Carroll SB (2005) More genes or more taxa? the relative contribution of gene number and taxon number to phylogenetic accuracy. Mol Biol Evol 22:1337–1344

Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804

SanMiguel PJ, Ramakrishna W, Bennetzen JL, Busso CS, Dubcovsky J (2002) Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome 5A(m). Funct Integr Genomics 2:70–80

Suoniemi A, Anamthawat-Jonsson K, Arna T, Schulman AH (1996) Retrotransposon BARE-1 is a major, dispersed component of the barley (*Hordeum vulgare* L.) genome. Plant Mol Biol 30:1321–1329

The Rice Chromosome 10 Sequencing Consortium (2003) In-depth view of structure, activity, and evolution of rice chromosome 10. Science 300:1566–1569

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796–815

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Tomkins J, Fregene M, Main D, Kim H, Wing R, Tohme J (2004) Bacterial artificial chromosome (BAC) library resource for positional cloning of pest and disease resistance genes in cassava (*Manihot esculenta* Crantz). Plant Mol Biol 56:555–561

Venter JC, Smith HO, Hood L (1996) A new strategy for genome sequencing. Nature 381:364–366

Vitte C, Bennetzen JL (2006) Eukaryotic transposable elements and genome evolution special feature: analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc Natl Acad Sci USA 103:17638–17643

Vitte C, Panaud O (2005) LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. Cytogenet Genome Res 110:91–107

Vogel J, Garvin D, Leong O, Hayden D (2006a) Agrobacterium-mediated transformation and inbred line development in the model grass *Brachypodium distachyon*. Plant Cell, Tissue and Organ Culture 84:100179–100191

Vogel JP, Gu YQ, Twigg P, Lazo GR, Laudencia-Chingcuanco D, Hayden DM, Donze TJ, Vivian LA, Stamova B, Coleman-Derr D (2006b) EST sequencing and phylogenetic analysis of the model grass *Brachypodium distachyon*. Theor Appl Genet 113:186–195

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

Wiedmann RT, Nonneman DJ, Keele JW (2006) Novel porcine repetitive elements. BMC Genomics 7:304