GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Transposable elements in *Phyllostachys pubescens* (Poaceae) genome survey sequences and the full-length cDNA sequences, and their association with simple-sequence repeats

M.B. Zhou, X.M. Liu, and D.Q. Tang

The Nurturing Station for the State Key Laboratory of Subtropical Silviculture,
Zhejiang A & F University, LinAn, Zhejiang Province, P.R. China

Corresponding author: D.Q. Tang
E-mail: tang@zafu.edu.cn

Genet. Mol. Res. 10 (4): 3026-3037 (2011)
Received January 5, 2011
Accepted March 31, 2011
Published December 6, 2011
DOI http://dx.doi.org/10.4238/2011.December.6.3

ABSTRACT. *Phyllostachys pubescens* is a woody bamboo with the highest ecological, economic, and cultural values of all bamboos in Asia. There is more genomic data available for *P. pubescens* than for any other bamboo species, including 2.12-Mb genome survey sequences (GSS) and 11.4-Mb full-length cDNA sequences (FL-cDNAs) currently deposited in GenBank. Analysis of these sequences revealed that transposable elements (TEs) are abundant, diverse and polyphyletic in the *P. pubescens* genome, of which Ty3-*gypsy* and Ty1-*copia* are the two most abundant families. Phylogenic analysis showed that both elements probably arose before the Bambusoideae separated from the other Poaceae subfamilies. We found evidence that the distribution of some intragenic TEs correlated with transcript profiles, of which *Mutator* elements preferred to insert in the transcripts of transcription factors. Additionally, we found that the abundance of SSRs in TEs (4.56%) was significantly higher than in GSS (0.098%) and in FL-cDNAs (2.60%) in *P. pubescen*s genome, and TA/AT and

CT/AG repeats were found to be intimately associated with *En/Spm* and *Mutator* elements, respectively. Our data provide a glimpse of the structure and evolution of *P. pubescens* genome, although large-scale sequencing of the genome would be required to fully understand the architecture of the *P. pubescens* genome.

Key words: *Phyllostachys pubescens*; Transposable elements;
Genome survey sequences; Full-length cDNA sequences;
Simple-sequence repeats

INTRODUCTION

Bamboo (family Poaceae, subfamily Bambusoideae) is a group of monocotyledonous plants divided into 77 genera and approximately 1030 species (Soderstrom and Ellis, 1987; Dransfield and Widjaja, 1995). Fifty genera and more than 500 species are found in China, of which *Phyllostachys pubescens* (synonym: *P. edulis*) is commercially the most important species, providing the third largest source of timber after Chinese red pine (*Pinus massoniana*) and China fir (*Cunninghamia lanceolata*). *P. pubescens* grows on 3 million ha (approximately 2% of the total forest area), an area that has doubled over the last 30 years (Fu, 2001).

Bamboo evolved from an ancestral grass and occupies an important phylogenetic node in the grass family (Clark, 1996; Klinkenborg, 2001). The genome sizes of bamboos are general large and were estimated to be between 2.45 and 5.3 pg DNA/2C, with temperate bamboo (*Phyllostachys*) falling within the range of 4.17-5.3 pg (Geilis et al., 1997). The genome size of *Olyra latiflia*, another herbaceous bamboo species, is approximately 9.5 pg and close to two times that of maize (Xu CM, Zhou MB, Dong WJ and Tang DQ, unpublished data). Given that amplification of transposable elements (TEs) is largely responsible for the big plant genome size (Bennetzen, 2002; Feschotte et al., 2002), it is reasonable to assume that large and diverse families of TEs will be found in bamboo genomes.

TEs are sequences of DNA that can move around to different positions within the genome of a single cell. Two broad classes of TEs are recognized based on their mechanism of transposition. Retrotransposons (class I) utilize an RNA intermediate and thus require reverse transcriptase to produce the DNA copy as well as an integrase for insertion into the host genome, whereas DNA transposons (class II) move directly as DNA and require a transposase to catalyze the necessary DNA cutting and joining reactions (Feschotte et al., 2002). TEs account for significant proportions of many eukaryotic genomes (SanMiguel and Bennetzen, 1998). For example, they account for at least 45% of the human genome (Lander et al., 2001) and 85% of the maize genome (Schnable et al., 2009). TEs are one of the propulsors of genome evolution. They cause both large-scale rearrangements and changes in the structure and expression of individual genes, through activities such as excision, integration, chromosome breakage, and ectopic recombination (Naito et al., 2009; Sinzelle et al., 2009). Many genes may have been assembled or amplified through the action of TEs, which can lead to infertility in heterozygous progeny. Therefore, TEs may be responsible for the rate at which such incompatibility is generated in separated populations (Bennetzen, 2002). TEs are also essential components of the transcriptome during the growth and development of the host organism (Lockton and

Gaut, 2009; Pritham, 2009). Transposase expression can be detected during tissue culture and growth under stress conditions, such as pathogen infection, pest infestation, drought, flooding, and exposure to radiation (Bennetzen, 2002; Jiao and Deng, 2007).

Compared to other members of the grass family, genomic data for bamboo are comparatively scarce, and no bamboo species genome has yet been sequenced on a large scale. Most of the available genome data are for *P. pubescens* (Tang, 2009). Presently, there are 2.12-Mb *P. pubescens* genome survey sequences (GSSs), including two BAC clones (GQ252886 and GQ252887, containing 113.2- and 139.3-kb genomic DNA, respectively), and 11.4-Mb 10608 full-length cDNA sequences (FL-cDNAs) deposited in GenBank. We compared the distribution of TEs between GSSs and FL-cDNAs, analyzed the phylogeny of Ty1-*copia* and Ty3-*gypsy* with the most prevalence, and characterized the transcript profiles of cDNA-TEs and searched for SSRs within TEs. These data provide important information on the structure and evolution of *P. pubescens* genomes and on the biology of TEs.

MATERIAL AND METHODS

Mining *P. pubescens* sequence data for TEs

Phyllostachys pubescens GSS and FL-cDNA data were downloaded from GenBank (http://www.ncbi.nlm.nih.gov/) on July 1, 2010. Redundant sequences were eliminated and overlapping sequences were spliced together using the CAP3 software (http://seq.cs.iastate.edu/cap3.html) (Huang and Madan, 1999). TEs were identified using RepeatMasker and RepeatProteinMask (http://www.repeatmasker.org) with rice (*Oryza sativa*) and maize (*Zea mays*) as the reference species. Open reading frames (ORFs) within FL-cDNAs were identified using ORF Finder (http://www.ncbi.nlm.nih.gov/projects/gorf/) and then conceptually translated into polypeptide sequences. Transcriptional factors were characterized by referring to the Plant Transcription Factor Database (http://planttfdb.cbi.pku.edu.cn/).

Phylogenetic analysis

The phylogenetic relationship among *P. pubescens* retrotransposons was determined by aligning reverse transcriptase amino acid sequences using CLUSTAL W (Thompson et al., 1994) with default parameters. Maize, rice and sorghum (*Sorghum bicolor*) retrotransposons from the Repbase Reports Database (http://www.girinst.org/repbase/) were used as reference sequences. Phylogenetic trees were constructed using the neighbor-joining (NJ), maximum parsimony (MP) and maximum likelihood (ML) methods with the PAUP software v4.0b10 (Swofford, 2002).

SSR detection

EST-trimmer (http://pgrc.ipk-gatersleben.de/misa/download/est_trimmer.pl) was used to remove terminal poly (A/T) runs from 5'- and 3'-ends until there were no occurrences of $(T)_5$ or $(A)_5$ within a 50-bp range. MISA (http://pgrc.ipk-gatersleben.de/misa/misa.html) was then used to search for SSRs within these FL-cDNAs, GSSs and TEs. The SSRs included mononucleotide repeats ≥10 bp in length, dinucleotide to hexanucleotide repeats with ≥6 repeat units, and interrupted composite SSRs with ≤100 bp of intervening DNA.

RESULTS

Distribution of TEs in *P. pubescens* GSSs and FL-cDNAs

After removal of redundant sequences, there remained 1.47-Mb non-redundant GSSs and 5.90-Mb non-redundant FL-cDNA sequences or contigs. They were analyzed by RepeatMasker to identify TEs, revealing 674 GSS-TEs (total 0.29 Mb, representing 20.42% of the GSS data) and 95 cDNA-TEs (total 13.52 kb, representing 0.32% of the FL-cDNA data).

The 674 GSS-TEs comprised 146 DNA transposons (total 0.04 Mb, representing 2.43% of the GSS data), 518 retrotransposons (total 0.25 Mb, 17.23% of the GSS data) and 10 uncharacterized TEs (total 5.31 kb, 0.36% GSS data). In contrast, the 95 cDNA-TEs comprised 54 DNA transposons (total 6.01 kb, 0.12% of the FL-cDNA data), 31 retrotransposons (total 5.87 kb, 0.12% of the FL-cDNA data) and 10 further uncharacterized TEs (total 1.63 kb, 0.03% of the FL-cDNA data). These results are summarized in Table 1.

Table 1. Transposable elements (TEs) in *Phyllostachys pubescens* genome survey sequences (GSSs) and full-length cDNA sequences (FL-cDNAs).

Names of TE families	Number of TEs in GSSs/FL-cDNAs		
	No.[a]	Length (bp)[b]	Proportion (%)[c]
RNA transposon	518/31	252790/5868	17.23/0.12
SINEs	0/0	0/0	0/0
Penelope	0/0	0/0	0/0
LINEs	11/1	3741/98	0.25/0
CRE/SLACS	0/0	0/0	0/0
L2/CR1/Rex	0/0	0/0	0/0
R1/LOA/Jockey	0/0	0/0	0/0
R2/R4/NeSL	0/0	0/0	0/0
RTE/Bov-B	0/0	0/0	0/0
L1/CIN4	11/1	3741/98	0.25/0
LTR elements	507/30	249049/5770	16.98/0.12
BEL/Pao	0	0	0
Ty1-*copia*	179/15	117876/3396	8.03/0.07
Ty3-*gypsy*	293/15	127712/2374	8.70/0.05
Retroviral	0	0	0
DNA transposons	146/54	35656/6014	2.43/0.12
Ac/Ds	45/6	12070/713	0.82/0.01
Tc1/marier	15/11	2157/1705	0.15/0.03
En/Spm	32/8	10606/620	0.72/0.01
Mutator	45/24	9922/2603	0.68/0.05
PiggyBac	0/0	0/0	0/0
Tourist/Harbinger/PIF	4/2	516/229	0.04/0
Uncharacterized TEs	10/10	5312/1634	0.36/0.03
Total		293758/13516	20.42/0.32

[a]Number of TEs in GSSs/FL-cDNAs. [b]The combined length in bp of all TEs in GSSs/FL-cDNAs. [c]Combined length of all TEs in GSSs/FL-cDNAs as a proportion of the combined length of non-redundant GSS/FL-cDNA sequence data.

Among the GSS DNA transposons, *Ac/Ds* elements were the most abundant (45 elements, covering 12.07 kb and 0.82% of the GSS data) followed by *Mutator* elements (also 45 elements, covering 9.92 kb and 0.68% of the GSS data). In contrast, among the cDNA DNA transposons, *Mutator* elements were the most abundant (24 elements, covering 2.60 kb and 0.05% of the FL-cDNA data) followed by the *Tc1*/*mariner* superfamily (11 elements, covering 1.71 kb and 0.03% of the FL-cDNA data). Among GSS retrotransposons, Ty3-*gypsy* elements were the most abundant (293 elements, covering 127.71 kb and 8.70% of the GSS data), followed by Ty1-*copia* elements (179 elements, covering 117.88 kb and 8.03% of the GSS data). The GSS retrotransposons were often found in clusters, e.g., in BAC GQ252887 there were 28 retrotransposons arranged in series. Compared with 179 examples (117.9 kb, 8.03%) and 293 (127.7 kb, 8.7%) in GSS, Ty1-*copia* and Ty3-*gypsy* elements were remarkably lower in FL-cDNAs, with only 15 examples of each element covering 2.37 kb (0.05%) and 3.40 kb (0.07%) of the FL-cDNA data, respectively. These results show that GSSs have a much higher TE content than FL-cDNAs, and Ty1-*copia* and Ty3-*gypsy* are the most abundant TEs in the *P. pubescens* genome (Table 1).

Evolution of Ty1-*copia* and Ty3-*gypsy* elements in *P. pubescens*

The evolution of *P. pubescens* retrotransposons was investigated by aligning the complete reverse transcriptase sequences from 24 Ty1-*copia* and 31 Ty3-*gypsy* elements with 36 elements from other Poaceae species (11 Ty1-*copia* and 10 Ty3-*gypsy* elements from rice, 3 Ty1-*copia* and 6 Ty3-*gypsy* elements from maize, and 6 Ty3-*gypsy* elements from sorghum). In the phylogenetic tree, we defined two retrotransposon clusters (*copia* and *gypsy*) as the largest and best-supported monophyletic groups (Figure 1). These groups were obtained regardless of the construction method (NJ, MP or ML). The Ty1-*copia* elements could be divided into four subclusters (I-IV) and the Ty3-*gypsy* elements into six subclusters (A-F). Every subcluster contained multiple retrotransposons, and all but one of the subclusters (the exception was *copia* cluster IV) contained retrotransposons from more than one species, indicating that all subclusters except *copia* IV predated the divergence of bamboo and the other grasses.

Analysis of the transcript profiles of cDNA-TEs

GenBank contains 10,608 *P. pubescens* FL-cDNAs, of which 4217 are expressed in leaf tissue, 3072 are expressed in embryos and 3318 are expressed in shoots (Peng et al., 2010). We investigated the distribution of the four most abundant TE families and found 24 *Mutator*, 11 *Tc1*/*mariner*, 15 Ty1-*copia*, and 15 Ty3-*gypsy* among these sequences. Among 24 *Mutator* elements, 10 were detected in leaf transcripts, seven in shoot transcripts and seven in embryo transcripts. Among 11 *Tc1/Mariner* elements, seven were detected in leaf transcripts, two in shoot transcripts and two in embryo transcripts. Among 15 Ty1-*copia* elements, eight were detected in leaf transcripts, three in shoot transcripts and four in embryo transcripts. Finally, among 15 Ty3-*gypsy* elements, seven were detected in leaf transcripts, three in shoot transcripts and five in embryo transcripts.

We also investigated the insertion sites of the 24 *Mutator* elements, noting that 19 elements had integrated into the 5' untranslated region (5'-UTR) and five into the coding region of the genes. A large proportion of the inserted genes encoded putative transcription factors (15 insertions); the others were involved in the energy cycle (four insertions), post-translational regulation (two insertions) and membrane transport (one insertion) (Table 2). The 14 genes encoding putative transcription factors included two regulated by hormones (FP099127 and FP099829) and three regulated by pathogens (FP091954, FP094062 and FP100486). The results show that *Mutator* elements have a strong preference for the 5'-UTRs of genes encoding transcription factors.

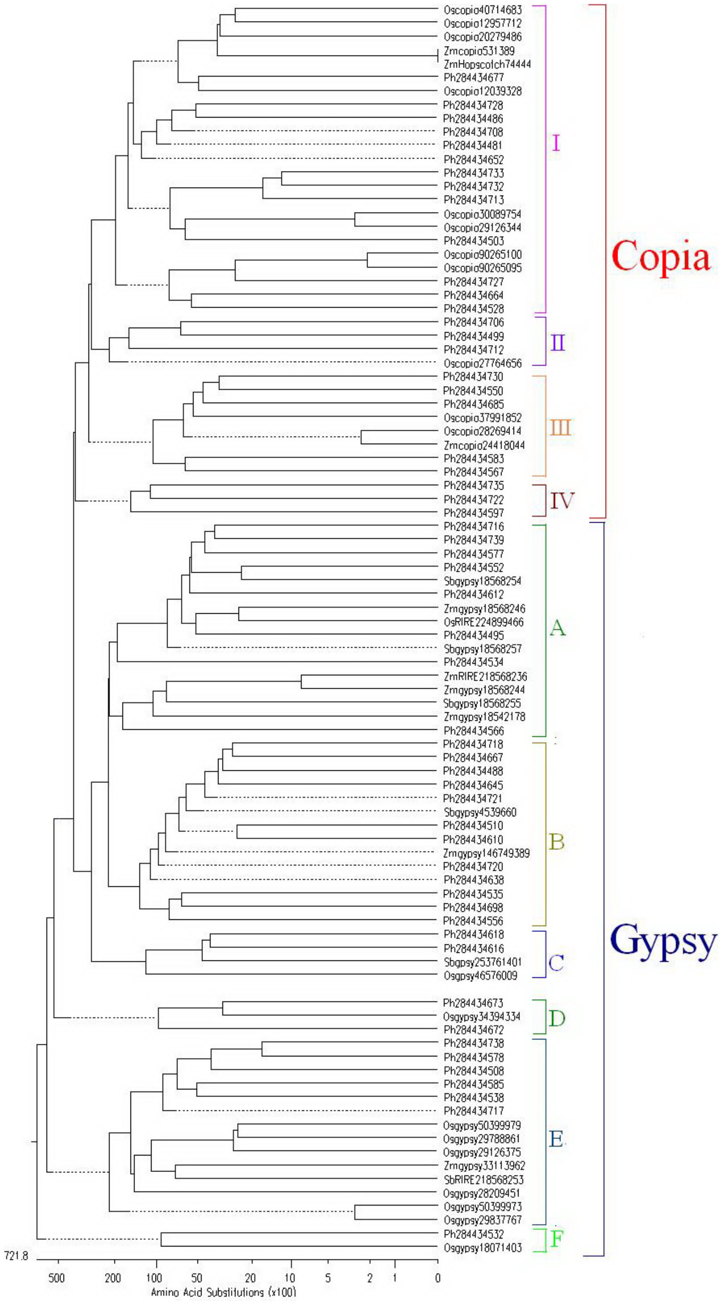

Figure 1. Phylogenetic analysis of Ty1-*copia* and Ty3-*gypsy* elements in *Phyllostachys pubescens*. Groupings defining lineages and sublineages of *P. pubescens* Ty1-*copia* and Ty3-*gypsy* elements are shown in different colors and named appropriately. The phylogenetic tree was generated by aligning 55 *P. pubescens* retrotransposons and 36 related elements from other Poaceae species: 11 Ty1-*copia* elements and 10 Ty3-*gypsy* elements from rice (Os), three Ty1-*copia* elements and six Ty3-*gypsy* elements from maize (Zm), and six Ty3-*gypsy* elements from sorghum (Sb).

Table 2. FL-cDNAs with integrated *Mutator* transposons.

GenBank No.	Insertion sites	Organ	Categories of genes	Homologous genes
FP091571	5'UTR	Leaf	Transcription factor gene	Homeobox transcription factor KNOX3 (*Hordeum vulgare*)
FP091749	5'UTR	Shoot		Unknown
FP091954	CDS	Embryo	Transcription factor gene	Pathogenesis-related transcriptional factor (Prb1) (*O. sativa*)
FP093400	5'UTR	Leaf	Cell energy cycle-related gene	ADP-ribosylation factor (*Zea mays*)
FP093768	CDS	Leaf	Transposase gene	Mutator-like transposase (*O. sativa*)
FP094062	CDS	Shoot	Transcription factor gene	Pathogenesis-related protein 1 precursor (PR-1) (*O. sativa*)
FP095802	5'UTR	Leaf	Transcription factor gene	Myb-like protein (*O. sativa*)
FP095913	5'UTR	Shoot	Transcription factor gene	Transcription factor MYBS2 (*O. sativa*)
FP096707	5'UTR	Shoot	Cell energy cycle-related gene	Proton-translocating NADH-quinone oxidoreductase (*Homalodisca coagulate*)
FP096801	5'UTR	Leaf	Post-translation-related gene	Serine/threonine protein kinases (*O. sativa*)
FP097565	5'UTR	Embryo	Transcription factor gene	GATA transcription factor (*Ricinus communis*)
FP097737	5'UTR	Leaf	Transcription factor gene	GlsA-related protein gene (*Chlamydomonas reinhardtii*)
FP099127	5'UTR	Embryo	Transcription factor gene	Auxin-responsive Aux/IAA family member (IAA15) (*Z. mays*)
FP099725	5'UTR	Shoot	Cell energy cycle-related gene	Adenine phosphoribosyltransferase (*O. sativa*) catalyzes the formation of AMP from adenine and 5-phosphoribosylpyrophosphate
FP099829	5'UTR	Embryo	Transcription factor gene	Auxin-responsive protein IAA7 (*O. sativa*)
FP100486	5'UTR	Shoot	Transcription factor gene	Pathogenesis-related transcriptional factor and ERF (*O. sativa*)
FP100707	CDS	Leaf	Transcription factor gene	TMPIT-like protein (*Sorghum bicolor*)
FP100818	5'UTR	Leaf	Membrane transport proteins	Plastid-lipid-associated protein 2 (*Arabidopsis thaliana*)
FP100934	CDS	Embryo	Post-translation-related gene	GPI-anchored protein (*O. sativa*)
FP100979	5'UTR	Leaf	Transcription factor gene	Transcription factor LcDREB3a (*Leymus chinensis*)
FP101158	5'UTR	Embryo	Transcription factor gene	CCAAT-binding transcription factor (*O. sativa*)
FP101391	5'UTR	Embryo	Transcription factor gene	PRLI-interacting factor G (*O. sativa*)
FP101428	5'UTR	Shoot	Cell energy cycle-related gene	GTP-binding protein sar1 (*Triticum aFL-cDNAivum*)
FP101553	5'UTR	Leaf	Transcription factor gene	Transcription activator/transcription factor (NAM, *Oryza sativa*)

UTR = untranslated region; CDS = coding sequence.

The distribution of SSRs in TEs

In many animals and plants, SSRs are distributed throughout the genome, but many numbers are located within TEs (Ramsay et al., 1999; Richard et al., 2008). We, therefore, investigated the distribution of SSRs among the 769 *P. pubescens* TEs and compared their abundance in TEs, GSSs and FL-cDNAs (Table 3). We found 69 SSRs in 63 TEs, covering

Table 3. Association between simple-sequence repeats (SSRs) and transposable elements (TEs) in the *Phyllostachys pubescens* genome.

Names of TE families	No.[a]	Length (bp)[b]	No. of SSR loci[c]	No. of TE-SSR sequences[d]	SSR proportion (%)[e]	No. of SSR motifs with different repeat units					
						Mono	Di	Tri	Tetra	Penta	Hexa
TEs	769	307274	69	63	4.56	12	49	6	1	1	0
En/Spm	40	11226	13	11	5.00	0	11	1	0	1	0
Mutator	69	12525	13	12	3.96	1	10	2	0	0	0
Ty1/*copia*	194	121272	2	2	0	0	1	1	0	0	0
Ty3/*gyspy*	308	130086	8	8	0.52	5	2	1	0	0	0
Other TEs	158	32165	33	30	2.41	6	25	1	1	0	0
GSSs/FL-cDNAs	907/7089	1467132/4942281	204/1614	111/1614[f]	0.098/2.60[g]	101/271	76/489	21/789	4/30	2/14	0/21

[a]Number of TE sequences identified in GSSs/FL-cDNAs. [b]The combined length of TEs identified in GSSs/FL-cDNAs. [c]Number of SSR loci identified in TEs. [d]Number of TEs that contains SSR loci. [e]SSR sequence length as a proportion (%) of TE sequence length. [f]Number of GSSs/FL-cDNAs that contains SSR loci. [g]SSR sequence length as a proportion (%) of GSS/FL-cDNA sequence length.

14,011 bp TE sequences (4.56% of the total TE sequences). In contrast, there were 204 SSRs in the non-redundant GSS data (1.47 Mb, 0.098% of the total GSS data) and 1614 SSRs in the non-redundant FL-cDNA data (4.90 Mb, 2.60% of the total FL-cDNA data). These results clearly show that *P. pubescens* SSRs are more abundant in TEs than in FL-cDNAs and in GSS.

Among the DNA transposons, SSRs were most likely to occur in *En/Spm* elements (5.00% of the total *En/Spm* DNA sequence), with 11 elements containing one SSR, and two containing two SSRs (Table 4). *Mutator* elements were the next most likely to contain SSRs, with 12 of 69 elements (3.96% of the total *Mutator* DNA sequences) containing at least one SSR, and one element containing two SSRs (Table 4). The situation was very different among the retrotransposons: close to 0% Ty1-*copia* DNA and only 0.52% Ty3-*gypsy* DNA sequences were made up of SSRs. Ten of the 12 SSR loci found in *En/Spm* elements were TA/AT repeats, and seven of the 13 SSR loci found in *Mutator* elements were CT/AG repeats, revealing a strong preference for dinucleotide repeat sequences. All 13 *Mutator* SSRs and 10 *En/Spm* SSRs were located in the 5'-UTR, revealing a strong preference for this location, at least in these specific transposon families.

Table 4. Distribution of SSRs in *En/Spm* and *Mutator* transposons.

ID	SSR motifs	Length (bp)	Starting	Ending	Location
SSR distribution in *En/Spm* transposons					
gi112350848	$(AT)_{17}$	34	9	42	5'UTR
gi284434591	$(AT)_{9}$	18	22	39	5'UTR
gi284434649	$(AT)_{6}$	12	19	30	5'UTR
gi284434671	$(AT)_{11}$-$(AT)_{15}$	87	28	114	5'UTR+CDS
gi284434701	$(TA)_{19}$	38	1	38	5'UTR
FP091991	$(GAGGA)_{6}$	30	109	138	CDS
FP091422	$(TA)_{22}(CA)_{9}$	62	12	73	5'UTR
FP097776	$(TA)_{23}$	46	1	46	5'UTR
FP100462	$(TA)_{31}$	62	14	75	5'UTR
FP100841	$(CGG)_{6}$	18	38	55	5'UTR
FP100858	$(AT)_{29}$	58	5	62	5'UTR
SSR distribution in *Mutator* transposons					
FP100733	$(TC)_{8}$-$(GGC)_{5}$	89	20	108	5'UTR
FP100664	$(AG)_{17}$	34	32	65	5'UTR
FP094905	$(CT)_{17}$	34	1	34	5'UTR
FP099988	$(CT)_{19}$	38	1	38	5'UTR
FP094782	$(CT)_{12}$	24	7	30	5'UTR
FP099842	$(CT)_{15}$	30	4	33	5'UTR
FP091749	$(CT)_{23}$	46	2	47	5'UTR
FP093400	$(GA)_{16}$	32	23	54	5'UTR
FP096707	$(GAA)_{8}$	24	36	59	5'UTR
FP096801	$(TC)_{8}$	16	2	17	5'UTR
FP099127	$(AG)_{18}$	36	40	75	5'UTR
FP099725	$(C)_{13}$	13	1	13	5'UTR

DISCUSSION

The previous phylogenetic studies of the grass family, based on a few chloroplast and nuclear genes, showed that Bambusoideae and Ehrhartoideae are sister groups and that there is a close relationship between bamboo and rice (Barker et al., 2001; Kellogg, 2001). Meanwhile TEs in the maize genome were identified to be the most abundant and diverse until now (Schnable et al., 2009). So rice and maize were selected as the reference species for identified *P. pubescens* TEs. Identified by RepeatMasker and RepeatProteinMask, just over 20% of the sequence was represented by TEs in 1.47-Mb non-redundant *P. pubescens* GSS data, which is similar to the 19.9% reported in rice (Turcotte et al., 2001), but is significantly lower than the 85% reported in maize (Schnable et al., 2009).

The *P. pubescens* genome is relative large (approximately 2034 Mb) and close to the maize genome in size, but more than five times larger than the genome of diploid cultivated rice (Gui et al., 2007). Given that the amplification of TEs is largely responsible for the large plant genome size (Bennetzen, 2002; Feschotte et al., 2002), there should be a higher TE content in *P. pubescens* GSSs than predicted in our study. One possible reason is that *P. pubescens* GSS data currently deposited in GenBank are insufficient (cover only 0.36% of the genome) to represent the whole *P. pubescens* genome. It is likely that the proportion of TEs in *P. pubescens* genome has been underestimated by focusing on publically available GSS data. One direct evidence is that our previous studies have shown that *mariner*-like, *PIF*-like and *Pong*-like elements are all very abundant in the *P. pubescens* genome, while relatively rare in the GSS data reported here (Zhong et al., 2010; Zhou et al., 2010a,b).

Compared to *P. pubescens* GSS data, the FL-cDNA data are abundant and likely to represent more than a quarter of bamboo genes and the third largest collection of FL-cDNA sequences next to those of *Arabidopsis* and rice. It provides the first large sequence dataset for studying the structure and function of a substantial portion of bamboo genes, and fills the gap in the grass family for comparative genomics (Peng et al., 2010). In our study, TEs in the transcriptome were stringently scanned. There are relatively few TEs within the FL-cDNA sequences compared to the genome average (TEs represent 0.32% of FL-cDNAs but >20% of GSSs), which indicates that most TEs are not expressed in *P. pubescens* transcriptome due to the tight regulation of TEs (Jiao and Deng, 2007). However, the presence of TEs in and around genes appears to be essential for the growth and development of the host organism (Lockton and Gaut, 2009; Pritham, 2009), because they are involved in the regulation of gene expression (Marino-Ramirez et al., 2005; Feschotte, 2008). In a seminal study, Jordan et al. (2003) reported that nearly 25% of experimentally characterized human promoters contain TE sequences, including empirically defined *cis*-regulatory elements. Furthermore, despite the strong conservation of gene expression patterns across different maize lines, *Mutator* transposition programmed by transcriptionally active *MuDR* can induce a 25% change in the anther transcriptome, reflecting widespread insertion of the *Mutator* transposon into genes encoding transcription factors (Skibbe et al., 2009). Our data show that 14 of 24 *Mutator* insertion sites in *P. pubescens* FL-cDNAs are located in genes encoding transcription factors, which indicates that *Mutator* transposons may influence host growth and development by influencing the regulatory factor of gene expression (Marino-Ramirez et al., 2005).

We previously reported a phylogenic analysis of the partial polypeptide sequences of 29 reverse transcriptase (approximately 90 amino acids) from Ty1-*copia* elements in *P. pubescens* and nine diverse cultivars, revealing that the elements were both diverse and abundant in the *P. pubescens* genome (Zhou et al., 2010c). Here we extended the analysis of the full-reverse transcriptase sequences from 24 Ty1-*copia* and 31 Ty3-*gypsy* elements from *P. pubescens* and 36 related elements from other Poaceae species (rice, maize and sorghum). This resulted in two major clusters, corresponding to the Ty1-*copia* and Ty3-*gypsy* families. These were further divided into 10 subclusters, all but one of which (*copia* subcluster IV) contained retrotransposon sequences from multiple species, indicating that most of the subclusters existed before divergence of *P. pubescens* and the other species of Poaceae (Figure 1). The available fossil evidence and the surviving basal lineages suggest that the Bambusoideae diverged from the rest of the Poaceae during the upper Cretaceous more than 65 Myrs ago (Guo and Li, 2002). It therefore appears that *P. pubescens* Ty1-*copia* and Ty3-*gypsy* elements are ancestral in origin, and originated more than 65 Myrs ago and have subsequently undergone extensive genetic and epigenetic diversification (Matsuoka and Tsunewaki, 1999).

SSRs are defined as DNA sequences 1-6 bp in length that are tandemly repeated a variable number of times. We also investigated the distribution of SSRs among *P. pubescens* TEs because previous reports have shown that many SSRs are located in TEs (Ramsay et al., 1999; Richard et al., 2008), e.g., 50% of human SSRs distributed within TEs (Scherer, 2008). As expected, the abundance of SSRs in *P. pubescens* TEs (4.56%) was significantly higher than in the genome (based on GSS analysis, 0.098%) and in expressed sequences (based on FL-cDNA analysis, 2.60%). Some studies have shown that some types of SSRs are intimately associated with some families of TEs, e.g., $(TA)_n$ dinucleotide repeats are frequently found in the 5'-UTR of the Micron (a microsatellite-targeting transposable element) in rice (Akagi et al., 2001; Temnykh et al., 2001). Similarly, TA/AT repeats and CT/AG repeats were found to be intimately associated with *En/Spm* and *Mutator* elements of *P. pubescens*, respectively (Table 3).

In conclusion, we investigated the distribution, diversity and evolution of TEs in the 7.37-Mb non-redundant *P. pubescens* sequence data available in GenBank, July 2010. TEs are relatively abundant, diverse and polyphyletic in the *P. pubescens* genome. Ty1-*copia* and Ty3-*gypsy* are the most abundant elements. These appear to predate the divergence of the Bambusoideae from the rest of the Poaceae, and have undergone significant genetic and epigenetic differentiation. We found evidence that the distribution of some intragenic TEs is correlated with transcript profiles, and we found that many *P. pubescens* TEs contain SSRs. Our data provide a tantalizing glimpse of the structure and evolution of *P. pubescens* genome, although large-scale sequencing of the *P. pubescens* genome would be required to fully understand the architecture of the *P. pubescens* genome.

ACKNOWLEDGMENTS

Research supported by special grants (#30771753 and #31070590) from the National Science Foundation of China and the Talents Program of the Natural Science Foundation of Zhejiang Province (#R303420).

REFERENCES

Akagi H, Yokozeki Y, Inagaki A, Mori K, et al. (2001). Micron, a microsatellite-targeting transposable element in the rice genome. *Mol. Genet. Genomics* 266: 471-480.

Barker NP, Clark LG, Davis JI, Duvall MR, et al. (2001). Phylogeny and subfamilial classification of the grasses (Poaceae). *Ann. Missouri Bot. Garden* 88: 373-457.

Bennetzen JL (2002). Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica* 115: 29-36.

Clark LG (1996). Diversity, Biogeography and Evolution of *Chusquea*. In: International Symposium on Bamboos (Chapman GP, ed.). Academic Press Ltd., London, 33-44.

Dransfield S and Widjaja EA (1995). Plant Resources of South-East Asia No. 7 Bamboos. Backhuys Publishers, Leiden.

Feschotte C (2008). TEs and the evolution of regulatory networks. *Nat. Rev. Genet.* 9: 397-405.

Feschotte C, Jiang N and Wessler SR (2002). Plant transposable elements: where genetics meets genomics. *Nat. Rev. Genet.* 3: 329-341.

Fu J (2001). Chinese Moso Bamboo: Its importance. *Bamboo* 22: 5-7.

Geilis J, Everaert I and De Loose M (1997). Genetic Variability and Relationships in *Phyllostachys* Using Random Amplified Polymorphic DNA. In: The Bamboos (Chapman GP, ed.). Linnean Society Symposium, Academic Press, London, 107-124.

Gui Y, Wang S, Quan L, Zhou C, et al. (2007). Genome size and sequence composition of moso bamboo: a comparative study. *Sci. China C Life Sci.* 50: 700-705.

Guo ZH and Li DZ (2002). Advances in the systematics and biogeography of the Bambusoideae (Gramineae) with remarks on some remaining problems. *Acta Bot. Yunnanica* 24: 431-438.

Huang X and Madan A (1999). CAP3: A DNA sequence assembly program. *Genome Res.* 9: 868-877.

Jiao Y and Deng XW (2007). A genome-wide transcriptional activity survey of rice transposable element-related genes. *Genome Biol.* 8: R28.

Jordan IK, Rogozin IB, Glazko GV and Koonin EV (2003). Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.* 19: 68-72.

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

Klinkenborg V (2001). Bamboo for Gardens. Times Book Rev, New York.

Lander ES, Linton LM, Birren B, Nusbaum C, et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.

Lockton S and Gaut BS (2009). The contribution of transposable elements to expressed coding sequence in *Arabidopsis thaliana*. *J. Mol. Evol.* 68: 80-89.

Marino-Ramirez L, Lewis KC, Landsman D and Jordan IK (2005). Transposable elements donate lineage-specific regulatory sequences to host genomes. *Cytogenet. Genome Res.* 110: 333-341.

Matsuoka Y and Tsunewaki K (1999). Evolutionary dynamics of Ty1-copia group retrotransposons in grass shown by reverse transcriptase domain analysis. *Mol. Biol. Evol.* 16: 208-217.

Naito K, Zhang F, Tsukiyama T, Saito H, et al. (2009). Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature* 461: 1130-1134.

Peng Z, Lu T, Li L, Liu X, et al. (2010). Genome-wide characterization of the biggest grass, bamboo, based on 10,608 putative full-length cDNA sequences. *BMC Plant Biol.* 10: 116.

Pritham EJ (2009). TEs and factors influencing their success in eukaryotes. *J. Heredity* 100: 648-655.

Ramsay L, Macaulay M, Cardle L, Morgante M, et al. (1999). Intimate association of microsatellite repeats with retrotransposons and other dispersed repetitive elements in barley. *Plant J.* 17: 415-425.

Richard GF, Kerrest A and Dujon B (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Micr. Mol. Bio. Rev.* 72: 686-727.

SanMiguel P and Bennetzen JL (1998). Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Ann. Bot.* 82: 37-44.

Scherer S (2008). A Short Guide to the Human Genome. Cold Spring Harbor University Press, Cold Spring, New York.

Schnable PS, Ware D, Fulton RS, Stein JC, et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* 326: 1112-1115.

Sinzelle L, Izsvak Z and Ivics Z (2009). Molecular domestication of transposable elements: from detrimental parasites to useful host genes. *Cell Mol. Life Sci.* 66: 1073-1093.

Skibbe DS, Fernandes JF, Medzihradszky KF, Burlingame AL, et al. (2009). Mutator transposon activity reprograms the transcriptomes and proteomes of developing maize anthers. *Plant J.* 59: 622-633.

Soderstrom TR and Ellis RP (1987). The Position of Bamboo Genera and Allies in a System of Grass Classification. In: Grass Systematics and Evolution (Soderstrom TR, Hilu KW, Campbell CS and Barkworth ME, eds.). Smithsonian Institution Press, New York, 225-238.

Swofford DL (2002). PAUP: Phylogenetic Analysis Using Parsimony (*and other methods). Version 4.0b 10. Sinauer Associates, Sunderland.

Tang DQ (2009). Genomic sequencing and its application for biological and evolutional research in bamboo. *Bamboo J.* 26: 1-10.

Temnykh S, DeClerck G, Lukashova A, Lipovich L, et al. (2001). Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res.* 11: 1441-1452.

Thompson JD, Higgins DG and 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.

Turcotte K, Srinivasan S and Bureau T (2001). Survey of transposable elements from rice genomic sequences. *Plant J.* 25: 169-179.

Zhong H, Zhou M, Xu C and Tang DQ (2010). Diversity and evolution of Pong-like elements in Bambusoideae subfamily. *Biochem. Syst. Ecol.* 38: 750-758.

Zhou MB, Lu JJ, Zhong H, Liu XM, et al. (2010a). Distribution and diversity of PIF-like transposable elements in the Bambusoideae subfamily. *Plant Sci.* 179: 257-266.

Zhou MB, Lu JJ, Zhong H, Tang KX, et al. (2010b). Distribution and polymorphism of mariner-like elements in the Bambusoideae subfamily. *Plant Syst. Evol.* 289: 1-11.

Zhou MB, Zhong H, Zhang QH, Tang KX, et al. (2010c). Diversity and evolution of Ty1-copia retroelements in representative tribes of Bambusoideae subfamily. *Genetica* 138: 861-868.