Bangladesh J. Bot. **50**(1): 69-77, 2021 (March)

GENOMIC CHARACTERISTICS OF CRITICALLY ENDANGERED *CHIMONOBAMBUSA HIRTINODA* C.S. CHAO & K.M. LAN

TENGFEI SHEN[1], YOUMIAO ZHENG[2], ZIMOU SUN[1] AND MENG XU[1*]

Co-Innovation Center for Sustainable Forestry in Southern China, Nanjing Forestry University, Nanjing 210037, China

Keywords: K-mer, Genome size, Heterozygous, Square bamboo

Abstract

With its unique square-shaped culm, *Chimonobambusa hirtinoda* C.S. Chao & K.M. Lan is a critically endangered species, and its natural habitat is solely restricted to Doupeng Mountain in Guizhou, China. Two small-insert libraries from *C. hirtinoda* were constructed and sequenced. Approximately 127.83 Gb of high-quality reads were generated and assembled into 9,320,997 contigs with a N50 length of 213bp, thereby producing 8,867,344 scaffolds with total length of 2.01 Gb. An estimated genome size of *C. hirtinode* was 2.86 Gb on the basis of k-mer frequency analysis, with the GC content of 45.40%. The repeat rate and heterozygous ratio were 74.11 and 1.48% in *C. hirtinoda* genome, respectively. Finally, 65,398 SSR loci were identified in the assembled contigs, including 58.66% tri-nucleotide, 27.42% di-nucleotide, 7.94% tetra-nucleotide, 3.67% penta-nucleotide, and 2.31% hexa-nucleotide. Results of this study are useful not only for ecological conservation of *C. hirtinoda*, but also for phylogenetic studies.

Introduction

As an important non-timber forest resource, Bambusoideae includes 115 genera and more than 1400 species, most of which are primarily distributed in Asia, South America and Africa. The subfamily Bambusoideae is the unique one that embraces woody members in the grass family (Poaceae) (Gu *et al*. 2016). Woody bamboos are of notable cultural and economic significance and have a long cultivation and utilization history, providing food and raw materials for construction and manufacturing all around the world. About 2.5 billion people rely on it economically and the annual trade value has exceeded 2.5 billion US dollars (Peng *et al*. 2013). The subfamily Bambusoideae in China comprises 37 genera with 500 species, many species play important roles in their ecosystems (Zhang *et al*. 2011). Because of their remarkable growth rate, infrequent reproduction and long flowering intervals, woody bamboos are very interesting but taxonomically challenging taxa (Zhang *et al*. 2011, Peng *et al*. 2013), with potentials for economic development and scientific research.

Chimonobambusa Makino (Bambusoideae, Poaceae), consisting of more than 30 taxa, are primarily distributed in China, Japan, Vietnam and Myanmar. *C. hirtinoda* belongs to Sect. *Oreocalamus* of the genus. Its internodes are usually slightly 4-angled, with a fulvous tomentose ring below each node. *C. hirtinoda* is a critically endangered species in China, and its natural distribution is solely restricted to Doupeng Mountain in Guizhou. Due to human interference activities such as bamboo shoot harvesting, the only *C. hirtinoda* population is declining (Su *et al*. 2016a,b), and this species is now listed as an IUCN critically endangered plant (http://www.incnredlist.org/search). So far, the study of *C. hirtinoda* remains as a blank sheet. With the fast development of next-generation sequencing (NGS) technology and the completion of the Moso bamboo (*Phyllostachys edulis*) whole-genome sequencing (Peng *et al*. 2013), increasing availability of genome and transcriptome data provide new insights on bamboo genetics and

*Author for correspondence: xum@njfu.edu.cn. [1]College of Forestry, Nanjing Forestry University, Nanjing 210037, China. [2]China National Bamboo Research Center, Hangzhou 310012, China.

evolution (Desai *et al*. 2015, Yeasmin *et al*. 2015, Sun *et al*. 2016). Here two small-insert libraries from *C. hirtinoda* were constructed and sequenced by using Illumina paired-end DNA sequencing technology. Furthermore, its genomic characteristics such as genome size, GC content and heterozygous rate, by analyzing k-mer frequency were revealed

Materials and Methods

The only population of *Chimonobambusa hirtinoda* survives in Doupeng Mountain Nature Reserve. Doupeng Mountain is located in the Miaoling mountainous region in the south of Guizhou Province, which is situated on the southeast slope of the Guizhou Plateau. It is also one of the watersheds and headstreams of the Pearl River and Yangtze River systems. The Nature Reserve has an area of 170 km^2 and belongs to a humid subtropical monsoon climate. The bamboo (*C. hirtinoda*) and broad-leaved tree mixed forest is mainly distributed between the east longitude of 107°31'05"" - 107°31'06" and the north latitude of 26°19'49" - 26°19'50", with an altitude of 1,080-1,190 m, a slope of 30°-32°, and an area of about 2,800 m^2. The accompanying species of *C. hirtinoda* are *Liquidambar formosana*, *Kalopanax septemlobus*, *Phoebe zhennan*, *Betule luminifera*, *Quercus fabri*, *Q. acutissima*, *Cinnamomum camphora*, *Rhus chinensis*, *Celtis sinensis*, *Rhododendron simsii*, *Lyonia ovalifolia*, and so on. There are a few vegetations under the forest.

Young leaves of *C. hirtinoda* were ground into powder in liquid nitrogen, and genomic DNA was isolated using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany). DNA was quantified using the ND-1000 spectrophotometer (Nanodrop, Wilmington, USA) and electrophoresis on 1% agarose gel. To prepare libraries for paired-end (PE) sequencing, all steps of the procedure were conducted according to the manufacturer's instructions (Illumina, USA), including fragmenting the genomic DNA, performing end repair, adding 'A' bases to the 3' end of the DNA fragments, ligating adapters to DNA fragments, purifying ligation products, enriching the adapter-modified DNA fragments by PCR, and validating the library. Finally, two PE genomic DNA libraries, with insert size of approximately 350 bp, were successfully constructed and further sequenced in the Illumina HiSeq 2500 sequencing platform.

In using CASAVA base-calling, the raw images were transformed into the FASTQ format of raw reads (raw data). Clean data (clean reads) were obtained by filtering reads containing adaptors, ploy-N larger than 10% reads, and low-quality reads with more than 50% $Q \leq 5$ bases from raw data. Furthermore, authors conducted some quality assessment methods, including base percentage composition, qualities distribution and error rate. All clean reads were used to estimating genome size, repetitive sequences, and heterozygosis. Based on k-mer analysis, information on peak depth and the number of 17-mers was revealed, then the genome size of C. *hirtinoda* was estimated by using the following algorithm: Genome size = k-mer num/peak depth (Varshney *et al*. 2011, Liu *et al.* 2013). The clean short reads were assembled *de novo* using SOAPdenovo software. Based on k-mer de bruijn graphs, all usable reads were realigned to the contig sequences, and then the PE relationship between reads was coincident between contigs. The scaffolds were constructed using insert size PEs. The GC content and average sequencing depth were estimated by the 10-kb non-overlapping sliding windows along the assembled sequence. Simple sequence repeats (SSRs) were identified in the genome sequence by using the MIcroSAtellite (MISA) search module. The parameters were set to detect perfect di-, tri-, tetra-, penta-, and hexa-nucleotide motifs with a minimum repeat length of 12 bp.

Results and Discussion

Illumina paired end (PE) sequencing generated 128.2 Gb of PE raw reads from the genomic DNA libraries of *Chimonobambusa hirtinoda.* All raw data were deposited to the NCBI SRA

database (accession no. SRP138012). For raw data, quality control analyses including the base percentage composition along reads, the distribution of base qualities and error rate along reads were performed for each library (Fig. 1), indicating high quality of sequencing data.

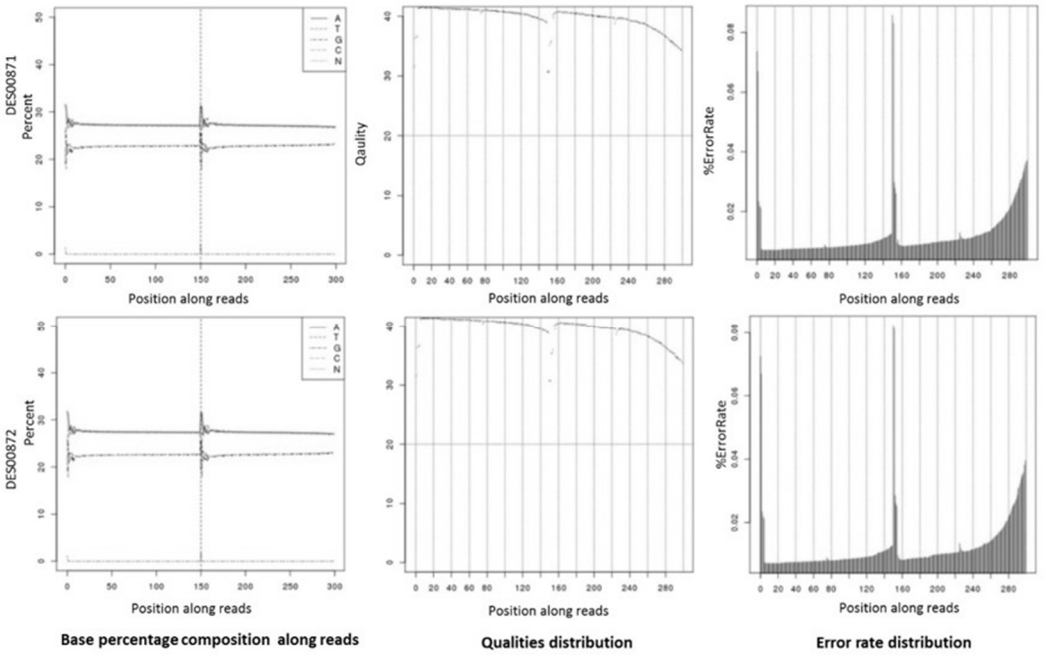

Fig. 1. Distribution of base composition along reads, base qualities and error rate along reads.

After filtering reads with adaptors or ambiguous nucleotides, and low-quality reads, a total of 127.83 Gb clean data were generated from the two PE libraries, and more than 97.6% of them had Phred quality scores at the Q20 level (Table 1). Further, 10,000 pairs of clean PE reads were randomly selected from each sequencing library and were used in BLAST searches against NCBI's NT database. BLAST analysis showed that the first five species of the top BLAST hit of 10,000 PE reads were closely related species of *C. hirtinoda* (Table 2) and provided a crude indication of no exogenous DNA pollution.

The genome size of *C. hirtinoda* was estimated by counting k-mer frequency of the127.83 Gb clean data (Table 1). Briefly, after optimization of k-mers, k-mer frequency counting was conducted, and the genome size can be estimated by using the following formulas: Genome size = k-mers num/peak depth, and revised genome size = genome size × (1-error rate). The peak depth was at 31× (Fig. 2), and the number of 17-mers was 90,031,861,576. The estimated genome size of 2904.25 Mb was revealed, and the revised genome size of *C. hirtinoda* was 2856.69 Mb (Table 3). Similarly, the minor peak at the position of the integer multiples of the main peak indicated about 74.11% of the repeat rate in *C. hirtinoda* genome, and the sub-peak at half of the main peak indicated about 1.48% of the heterozygosis rate in this genome (Fig. 2).

A total of 127.83 Gb clean reads were generated from the two small-insert libraries, about 44.75× coverage of the estimated 2,856.69 Mb, much greater than 30× coverage, which was required for successful assembly. Assembly with k-mer 41 by SOARdenovo produced a contig with an N50 of ~213 bp, a longest contig length of ~18.88 kb, and a total length of ~1.95 Gb (Table 4). A sequence with a scaffold N50 length of ~257 bp, a total length of ~2.01 Gb, and a

Table 1. Summary of sequencing quality assessment.

Library ID	Insert size	Raw reads	Clean Bases	Error (%)	Q20 (%)	Q30 (%)	GC (%)
DES00871	350 bp	222, 808, 123	66.65G	0.02	97.76	94.76	45.36
DES00872	350 bp	204, 547, 741	61.18G	0.02	97.60	94.46	45.42

Table 2. The first five species of top BLAST hit of 10,000 PE reads.

	Species	L number (%)	R_number (%)	Total (%)
DES00871	*Phyllostachys edulis*	498(9.96)	414(8.28)	9.12
	Oryza sativa	20(0.4)	16(0.32)	0.36
	Ferrocalamus remosivaginus	18(0.36)	13(0.26)	0.31
	Arundinaria gegantea	13(0.26)	15(0.30)	0.28
	Brachypodium distachyon	6(0.12)	12(0.24)	0.18
DES00872	*Phyllostachys edulis*	499(9.98)	448(8.96)	9.47
	Oryza sativa	10(0.20)	21(0.42)	0.31
	Arundinaria gegantea	12(0.24)	11(0.22)	0.23
	Brachypodium distachyon	14(0.28)	6(0.12)	0.20
	Ferrocalamus remosivaginus	10(0.20)	9(0.18)	0.19

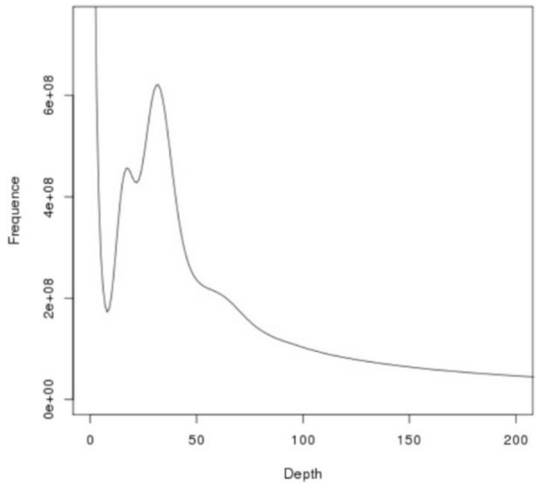

Fig. 2. K-mer (k-17) analysis for estimating the genome size of *C. hirtinoda.*

longest scaffold length of ~24.1 kb (Table 4). For these assembled contigs longer than 500 bp, authors further analyzed the relationship between contig size/number and coverage depth (Fig. 3). As shown in Fig. 3, two distinct peaks appeared at around ~23× the coverage depth and at ~11× the coverage depth, respectively. The difference in read coverage of the contigs may be derived from heterozygosity in the genome, indicating the *C. hirtinoda* genome with a higher heterozygosity rate.

GC content was an important factor contributing to sequence bias on the Illumina sequencing platform, too high and too low GC contents may seriously affect genome assembly. To estimate sequencing bias in more detail, 10-kb non-overlapping sliding windows along assembled sequence was used to compare GC content and average sequencing depth (Fig. 4). The X-axis was GC content percent across every 10-kb non-overlapping sliding window. The Y-axis was coverage sequencing depth. The right slide was coverage depth, and the top part was GC content distribution. The density points (red scatter plot) only concentrated within the 40-60% range. As shown in Fig. 2, a homozygous peaks appeared at ~23× the coverage depth from the right slide, and the main peak of GC content distribution was at around GC content of ~ 45.40% from the top part, which was identical with the GC content of the clean data.

Table 3. Estimation of the genome size based on k-mer statistics.

K-mer	K-mer number	K-mer depth	Genome size (Mb)	Revised genome size (Mb)	Heterozygous Ratio (%)	Repeat (%)
17	90, 031, 861, 576	31	2904.25	2856.69	1.48	74.11

Table 4. Statistics of the genome assembly based on 41-mer de bruijn graphs.

	Total length (bp)	Total number	Longest (bp)	N50 (bp)	N90 (bp)
Contig	1, 954, 261, 466	9, 320, 997	18,884	213	112
Scaffold	2, 011, 570, 804	8, 867, 344	24,097	257	114

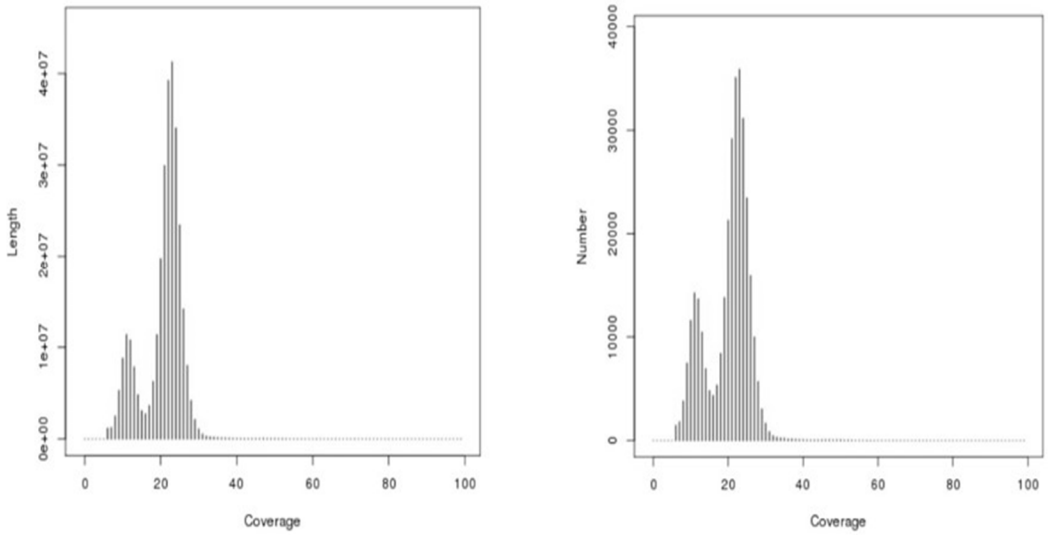

Fig. 3. Distribution of the contig size/number (Y-axis) and their coverage depth (X-axis).

The MISA module was used to detect for simple sequence repeat (SSR) loci, and only SSR were considered to contain di-, tri-, tetra-, penta- and hexa-nucleotide motifs with a minimum repeat length of 12bp. A total of 65,398 SSRs were identified in the assembled contigs (Fig. 5A), including 58.66% tri-nucleotide (38,364), 27.42% di-nucleotide (17,929), 7.94% tetra-nucleotide

(5,195), 3.67% penta-nucleotide (2,401), and 2.31% hexa-nucleotide (1,509). Within the dinucleotide repeat motifs (Fig. 5B), the AG/CT was the most abundant, accounting for 66.91%, followed by AC/TG (14.66%). Among the trinucleotide repeat motifs (Fig. 5C), the dominant motifs were GCC/GGC and TCC/GGA, accounting for 24.66 and 19.10%, respectively. The distribution of SSR motif length was summarized by counting the numbers of SSR per length, which ranged in length from 12 to 120 bp (Table 5).

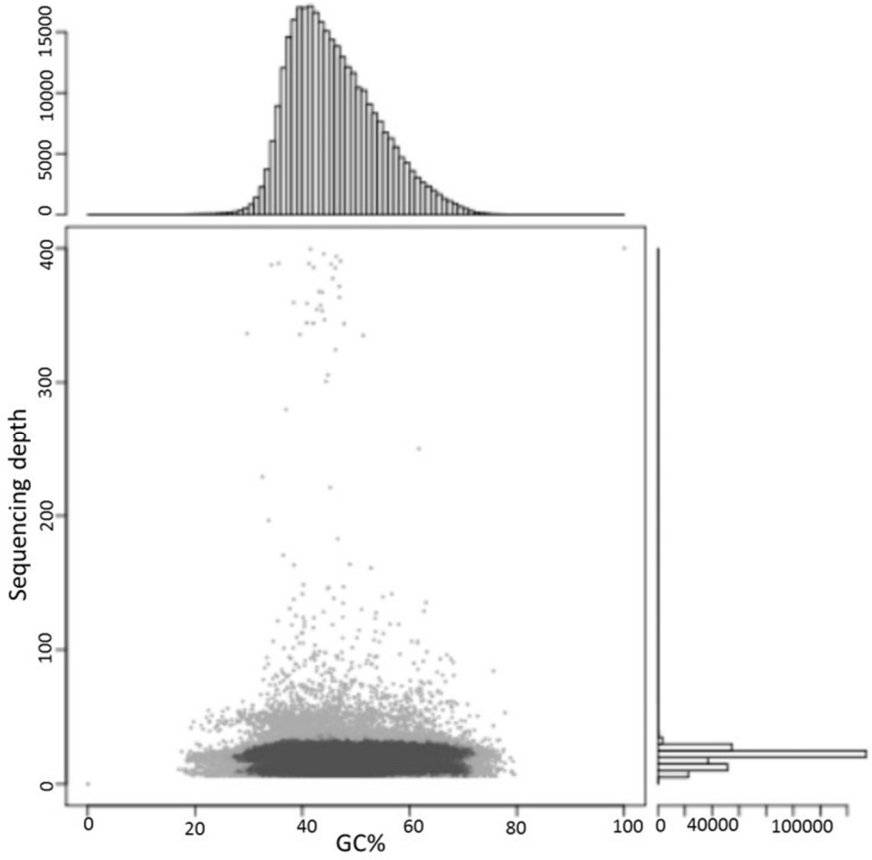

Fig. 4. GC content and sequencing depth of the of *C. hirtinoda* genome data used for assembly. The X-axis was GC content per cent across every 10 kb non-overlapping sliding window. The Y-axis was coverage sequencing depth. The right slide was coverage depth, and the top part was GC content distribution.

Genome size is a key biodiversity character for evolution and speciation, and it is the most fundamental genetic property of organisms. Estimation of genome size can be used for phylogenetic study, intergeneric classification, taxa delimitation and hybrid identification (Guo *et al*. 2015). Flow cytometry as a traditional experimental method has been used extensively used for estimation of nuclear genome size in plants. With the fast development of high-throughput sequencing technologies, the k-mer frequency analysis of sequenced reads has been regarded as a general and assembly-independent method for estimating genomic characteristic such as genome size, repeat structure and heterozygous rate (Liu *et al*. 2013).

High-throughput sequencing platforms can easily generate a higher coverage (>30×) of read data, which makes the k-mer analysis more accurate compared to the low-coverage (<10) required by traditional Sanger sequencers (Liu *et al*. 2013). Clean reads of 127.83 Gb were generated using Illumina HiSeq 2500 sequencer with a PE pattern, about 44.75× coverage of the estimated 2,856.69 Mb, much greater than 30× coverage. In the Bambusoideae, only the moso bamboo genome has been sequenced. The genome size of moso bamboo was 2.05 Gb, representing 95% of the genomic region, while its estimated size was approximately 2.07 to 2.10 Gb, which was supported by the analysis of the distribution of 51-mer frequencies (Peng *et al*. 2013). From the present genome survey data, the estimated genome size of *C. hirtinoda* was 2.86 Gb using the clean data for k-mer analysis, which was almost four times of *Sorghum bicolor* (Paterson *et al*. 2009) and six times of *Oryza sativa* (Yu *et al*. 2002).

Table 5. SSR motif length distribution in *C. hirtinoda*.

12 ≤ Length < 30	No.	30 ≤ Length < 50	No.	50 ≤ Length < 70	No.	Length ≥ 70	N0.
Length = 12	36288	Length = 30	730	Length = 50	36	Length = 70	5
Length = 14	3571	Length = 32	248	Length = 51	4	Length = 72	10
Length = 15	6028	Length = 33	28	Length = 52	43	Length = 74	2
Length = 16	6352	Length = 34	158	Length = 54	38	Length = 75	1
Length = 18	3190	Length = 35	33	Length = 55	4	Length = 76	2
Length = 20	3540	Length = 36	191	Length = 56	13	Length = 78	3
Length = 21	641	Length = 38	75	Length = 57	4	Length = 80	2
Length = 22	585	Length = 39	11	Length = 58	18	Length = 82	2
Length = 24	1986	Length = 40	70	Length = 60	23	Length = 84	1
Length = 25	393	Length = 42	69	Length = 62	8	Length = 86	1
Length = 26	302	Length = 44	60	Length = 63	3	Length = 90	1
Length = 27	137	Length = 45	10	Length = 64	5	Length = 98	1
Length = 28	330	Length = 46	52	Length = 65	1	Length = 102	1
		Length = 48	67	Length = 66	11	Length = 110	1
				Length = 68	5	Length = 111	1
				Length = 69	1	Length = 114	1
						Length = 120	1
Total (65398)	63,343		1,802		217		36

Guanine-cytosine (GC) content is the percentage of nitrogenous bases on a DNA or RNA molecule, or that of the whole genome. The GC content percentage as well as GC ratio within a genome is markedly variable, and coding sequences are often characterized by having a higher GC content in contrast to that of the entire genome. GC content is also variable with different organisms, the process of which is envisaged to be contributed to by variation in selection, mutational bias, and biased recombination-associated DNA repair (Birdsell *et al*. 2002). The GC content of *Populus trichocarpa* is 34.15% (Tuskan *et al*. 2006), and that of another common model organism, *Arabidopsis thaliana*, is 36.7% (Initiative *et al*. 2000). The *C. hirtinoda* genome had a mid-GC content of 45.4%, which was lower than that of *Z. mays* (46.91%) (Schnable *et al*.

2009), while higher than that of sorghum (44.16%) (Paterson *et al*. 2009), rice (43.73%) (Yu *et al*. 2002) and moso bamboo (43.9%) (Peng *et al*. 2013).

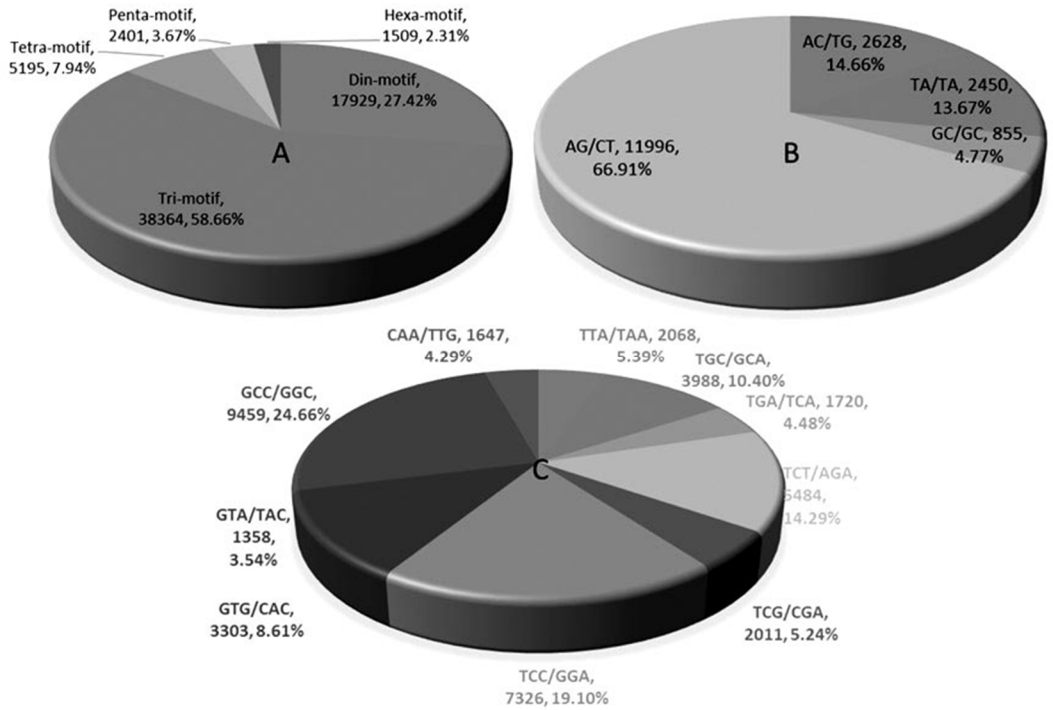

Fig. 5. SSR motif type and frequency in *C. hirtinoda*. (A) Frequency of di-, tri-, tetra-, penta- and hexa-nucleotide motifs. (B) Frequency of different di-nucleotide SSR motifs. (C) Frequency of different tri-nucleotide SSR motifs.

In conclusion, this is the first report of genomic characterization within this taxon. The estimated genome size of *C. hirtinoda* was 2.86 Gb, with a mid-GC content of 45.40%. The repeat rate and heterozygous ratio were 74.11 and 1.48%, respectively.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2016YFC0502605), the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

References

Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C and Gnirke A 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biology **12**(2): R18.

Birdsell JA 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Molecular Biology & Evolution **19**(7): 1181-1197.

Cheung MS, Down TA, Latorre I and Ahringer J 2011. Systematic bias in high-throughput sequencing data and its correction by BEADS. Nucleic Acids Research **39**(15): e103.

Desai P, Gajera B, Mankad M, Shah S, Patel A, Patil G, Narayanan S and Kumar N 2015. Comparative assessment of genetic diversity among Indian bamboo genotypes using RAPD and ISSR markers. Molecul. Biol. Reports **42**(8): 1265-1273.

Gu YS, Liu HY, Wang HL, Li RC and Yu JX 2016. Phytoliths as a method of identification for three genera of woody bamboos (Bambusoideae) in tropical southwest China. J. Archaeol. Sci. **68**: 46-53.

Guo LT, Wang LS, Wu QJ, Zhou XG and Xie W 2015. Flow cytometry and K-mer analysis estimates of the genome sizes of *Bemisia tabaci* B and Q (Hemiptera: Aleyrodidae). Frontiers in Physiology **68**: 46-53.

Initiative TAG 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana.* Nature **408**:796-815.

Liu BH, Shi YJ, Yuan JY, Hu XS, Zhang H, Li N, Li ZY, Chen YX, Mu DS and Fan W 2013. Estimation of genomic characteristics by analyzing k-mer frequency in *de novo* genome projects. Quantitative Biology **6**:144.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer H, Hellsten U, Mitros T, Poliakov A, *et al*. 2009. The *Sorghum bicolor* genome and the diversification of grasses. Nature **457**:551-556.

Peng Z, Lu Y, Li L, Zhao Q, Feng Q, Gao Z, Lu H, Hu T, Yao N, Liu K, *et al*. 2013. The draft genome of the fast-growing non-timber forest species moso bamboo (*Phyllostachys heterocycla*). Nature Genetics **45**(4): 456-461, 461e451-452.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, Liang CZ, Zhang JW, Fulton L, Graves TA, *et al*. 2009. The B73 Maize genome: Complexity, diversity, and dynamics. Science **326**(5956):1112-1115.

Sun H, Li L, Lou Y, Zhao H and Gao Z 2016. Genome-wide identification and characterization of aquaporin gene family in moso bamboo (*Phyllostachys edulis*). Molecular Biology Reports **43**(5): 437-450.

Su C, Zhu S, Li T, Zhou X and Wan N 2016a. Clone population structure and dynamic of *Chimonobambusa hirtinoda* under human disturbance. J. Fujian Forest. Sci. Technol. **43**(3): 62-66.

Su C, Zhu S, Zhang H and Luo G 2016b. Research on rhythm of shooting and growth for critically endangered *Chimonobambusa hirtinoda*. J. Fujian Forest. Sci. Technol. **43**(3): 153-156.

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, *et al*. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science **313**: 1596.

Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MT, Azam S, Fan G, Whaley AM, *et al*. 2011. Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nature Biotechnology **30**(1): 83-89.

Yeasmin L, Ali MN, Gantait S and Chakraborty S 2015. Bamboo: An overview on its genetic diversity and characterization. J. Biotech **5**(1): 1-11.

Yu J, Hu SN, Wang J, Wong GKS, Li SG, Liu B, Deng YJ, Dai L, Zhou Y, Zhang XQ, *et al*. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science **313**:1596.

Zhang YJ, Ma PF and Li DZ 2011. High-throughput sequencing of six bamboo chloroplast genomes: Phylogenetic implications for temperate woody bamboos (Poaceae: Bambusoideae). PLoS One **6**(5): e20596.

(Manuscript received on 17 July, 2019; revised on 30 March, 2020)