Genomic Resources of Broomcorn Millet: Demonstration and Application of a High-throughput BAC Mapping Pipeline

Wei Xu[1], Mengjie Liang[1], Hao Wang[1], Meizhong Luo[1*]

[1]College of Life Science and Technology, Huazhong Agricultural University, Wuhan 430070, China

*Correspondence: mzluo@mail.hzau.edu.cn

Equal contribution: Wei Xu and Mengjie Liang

Abstract

Background

With high-efficient water-use and drought tolerance, broomcorn millet has emerged as candidate for food security. To promote its research process for molecular breeding and functional research, a comprehensive genome resources is of great important.

Results

Herein, we constructed the first BAC library for broomcorn millet, generated BAC end sequences and integrated BAC clones into genome by a novel pipeline for BAC end profiling depending on clone-array pooled shotgun sequencing strategy and Illumina sequencing technology. The BAC library is consisted of 76,023 clones with average insert length of 123.48 Kb, about 9.9x coverage of 850 Mb genome. Then, 8262 of 9216 clones were mapped on broomcorn millet cultivar longmi4 genome using our pipeline. Furthermore, we also extracted and assembled unmapped reads against longmi4 genome. A total of 135 deletion sequences, 64 specific sequences and some sample contamination sequences were identified.

Conclusions

BAC clones in this library can be browsed and obtained from our website (http://flyinguineapig.com/gresource/JBrowse-1.16.5/index.html). This pipeline designed for BAC end profiling can greatly reduce the cost of acquiring BAC end sequences and shorten the period of the experiment compared with the Sanger sequencing method. These high-quality BAC clones related with genome in this study provide a useful and convinient genomic

resource for genome gap filling, complex segment sequencing, FISH, functional research, and genetic engineering.

1 Introduction

With the increased global water scarcity caused by climate change and population growth, it is of great importance to exploit the high-efficient water-use crop for the food security of the human in the future. Broomcorn millet (*Panicum miliaceum*), also known as proso millet, panic millet, and wild millet, is one of the traditional five-grain crops in the north of China [1]. It is a typical C4 plant with high photosynthetic efficiency. It also has a high-efficient utilization ratio of water resource and a capacity for drought resistance, exquisitely adapting to semi-drought or drought conditions [2]. Furthermore, its growing cycle, 60–90 days from sowing to maturity, is shorter than other cereals [3]. Broomcorn millet contains more protein than most grains, and a relatively balance array of trace elements and vitamins. More than 8700 accessions of broomcorn millet including landraces and cultivars are conserved in the National Gene Bank of the Institute of Crop Science, Chinese Academy of Agricultural Sciences, thus providing an abundance of resources for genetic improvement of broomcorn millet.

Broomcorn millet is an allotetraploid (2n = 4x = 36). At present, genome sequences of two broomcorn millet varieties decoded by Chinese researchers are available, and both of them have a high-quality chromosome-scale assembly [4, 5]. These genome sequences provide a foundation for molecular breeding of broomcorn millet. However, many gaps and unplaced contigs still exist and a perfect genome should be pursued.

…….

Hence, we developed a new method for decoding BAC end seqeunces efficiently. This method was tested on the BAC library of broomcorn millet. Those clones with assigned BESs were further associated with the longmi 4 genome to facilitate pick. In addition, the species-specific sequences were found by unmapped reads.

2 Results

2.1 BAC library construction

A BAC library of broomcorn millet was constructed with the restriction enzyme *Hin*dIII using high-molecular-weight genomic DNA prepared from etiolated seedlings. In total, the library consists of 76,032 BAC clones, which were arrayed into 198 independent 384-well microtiter plates. Insert sizing of randomly picked clones showed that the majority of genomic BAC inserts fell into the length range of 97–145.5 kb with an average insert size of 132 kb (Fig. 1).

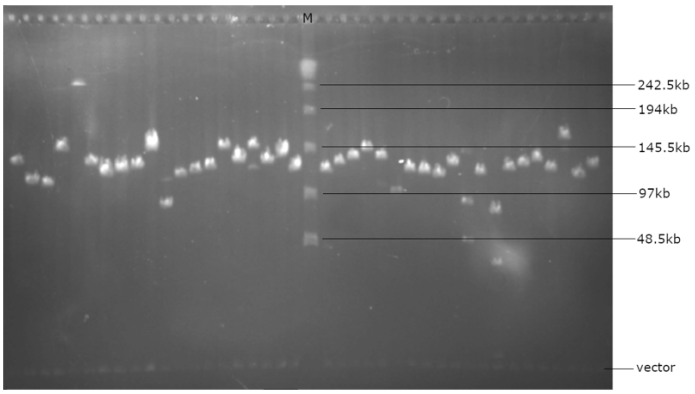

Fig. 1 The insert size of randomly selected BAC clones determined by PFGE. The maker in the middle is λ DNA ladder.

2.2 Construction and sequencing of DNA pools

In order to obtain BAC end sequences with high efficiency and low cost, we designed and performed a pipeline based on the clone array pooled shotgun sequencing strategy (Fig. 2A). Randomly selected 24 384-plates (each plate consists of 16 x 24 clones) from the broomcorn millet BAC library were arranged to a square superpool with 6 row plates and 4 column plates as shown in Fig. 2B. Therefore, the superpool consisted of 96 row pools (6 plates x 16) and 96 column pools (4 plates x 24). These row pools and column pools were called as the secondary pools, and each secondary pool also consisted of 96 BAC clones. In total, one superpool consisted of 192 secondary pools and 9216 individual BAC clones (Fig. 2B).

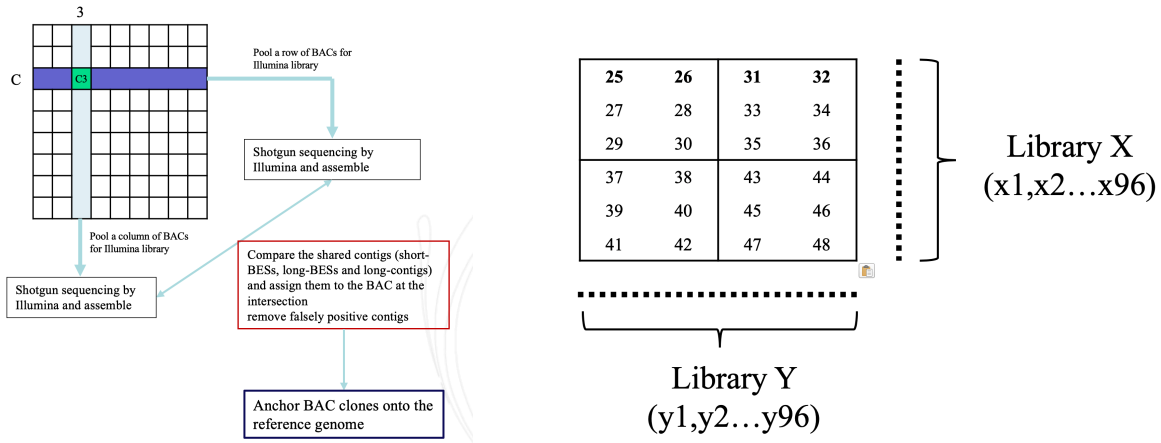

Fig. 2 The strategy of BAC end sequence localization based on CAPSS and the layout of 24 384-plates.

DNAs of the secondary pools were extracted and Illumina sequencing libraries with individual index sequences for the secondary pools were prepared. Finally, 96 row pool libraries were mixed in equal amounts as the library X, and 96 column pool libraries were also mixed in equal amounts as the library Y. The average insert lengths of library X and library Y analyzed using an Agilent 2100 Bioanalyzer (Fig. S1) are 491 bp and 443 bp, respectively. Their peaks are single and the fragment sizes are suitable for sequencing. Illumina sequencing with paired-end protocol (PE150) using the Illumina HiSeq 2000 sequencing platform was accomplished. Library X generated 130.82 Gb paired end reads, and library Y generated 133.50 Gb paired end reads. Trimmomatic was employed for trimming adaptors and filtering low-quality or shorter reads. FastQC was employed for checking the quality of the preprocessing reads. The final base qualities were substantially greater than Q28, which means that the base correction rate is greater than 99.84% (Q = -10 log10Pe). Next, the valid reads were obtained by filtering the reads traced to *E. coli* DH10B reference genome using Bowtie2. Finally, we obtained valid yields of libraries X and Y of 60.40 Gb and 102.37 Gb, respectively (Table 1). Demultiplexer was employed for demultiplexing and generating each secondary pool reads according to the 7-bp index (Supp. Table 1). The average valid sequencing depths of row pools and column pools were 45x and 78x, respectively (Fig. S4).

Table 1 The summary of NGS data.

Library	Total size (Gb)	Size after QC (Gb)	Valid size (Gb)
X	130.82	103.78	60.40
Y	133.50	107.92	102.37

2.3 Parsing BAC end sequences

A pipeline was designed as shown in Fig. 3. On the one hand, the paired-end reads overlapping with the border sequences of the vector harboring the *Hin*dIII restriction enzyme site were extracted from the valid data of each secondary pool, and then assembled using Cap3 based on the overlap-layout-consensus method. These short BAC end sequences in the secondary pools were assigned into the corresponding wells by clone array pooled shotgun sequencing (abbr. CAPSS) strategy. The vector sequence parts in the consensus were truncated to obtain short BAC end sequences (abbr. BES) starting with AAGCTT (*Hin*dIII site). Table S2 summarized the number of BAC end sequences in detail. There were 8183 (88.79%) and 7897 (85.69) wells to all 9216 wells that had only one forward or reverse BES respectively. All delivered BAC end sequences had an average length of 358 bp, which was consistent with the insert size of the Illumina library (Fig. S2).

On the other hand, the valid NGS data of each secondary pool were assembled using SPAdes, and the contig N50 sizes of all row pools and column pools were counted (Fig. S3). The average N50 of all row pools and all column pools were 13.97 kb and 11.45 kb, respectively. Likewise, the vector sequence parts at the ends of the contigs were removed to retain long BAC end sequences starting with AAGCTT. Finally, these long BAC end sequences of the secondary pools were assigned into the corresponding wells in 384-plates. Table S3 summarized the number of BAC end sequences in detail. There were 5454 (59.18%) and 5108 (55.43%) wells to all 9216 wells that had only one forward or reverse BAC end sequences respectively. The N50 of the long BAC end sequences was 60.87 kb (Fig. S4).

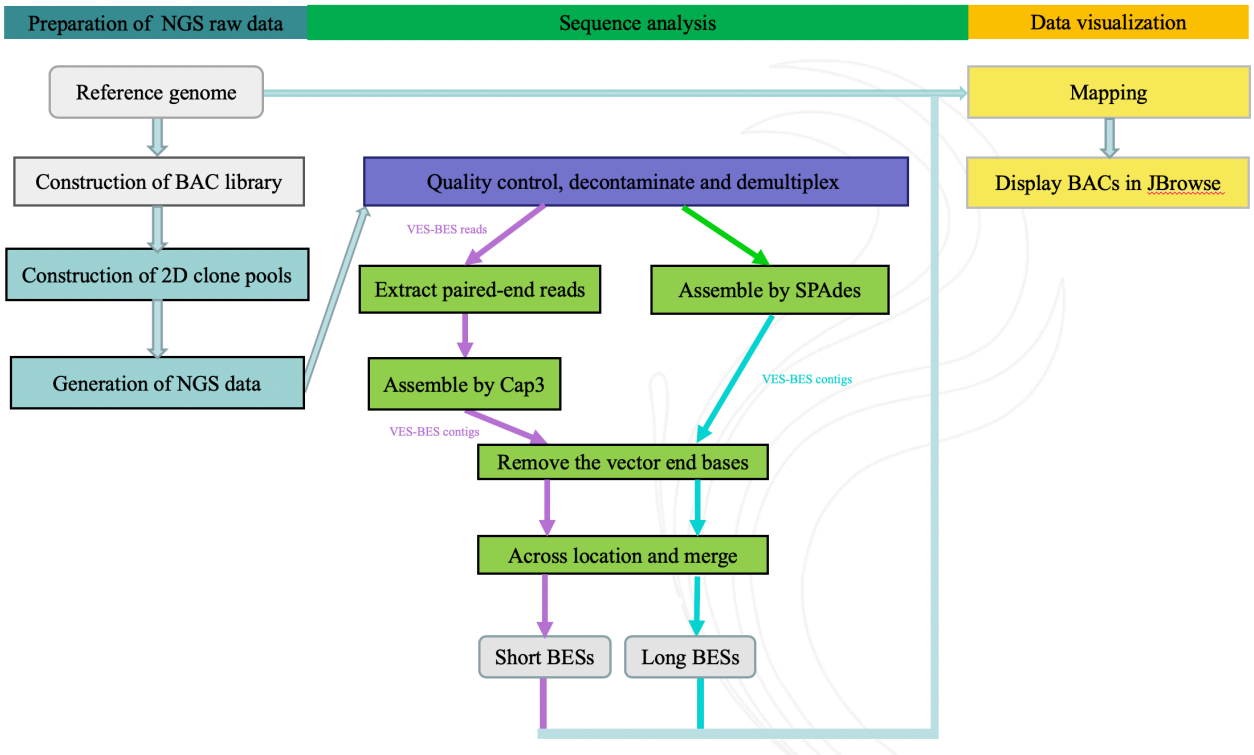

Fig. 3 The pipeline of BAC end sequence analysis.

2.4 Determination of BAC locations on the genome

In order to determine the locations of the broomcorn millet BAC clones on the genome using BAC end sequences, we aligned the short and long BAC end sequences onto the cultivar longmi4 genome with Blastn. Table 2 listed the alignment information in detail. In the alignment results of BAC end sequences with repeats, there were 14862 short BAC end sequences mapped to the genome that consist of 12971 unique hitted sequences and 1891 multiplex hitted sequences , and correspondingly there were 12120 long BAC end sequences uniquely hitted in the genome. In the alignment results of the BAC end sequences without repeats, there were 7760 short BAC end sequences mapped to the genome that consist of 7295 unique hitted sequences and 465mulitiplex hitted sequences, and conrespondingly there were 10626 long BAC end sequences uniquely hitted in the genome. In order to map as many as possible BACs onto the reference genome, the alignment results of the end sequences with repeats were adopted. We wrote a python script to extract the blast results from short and long BESs. As a result, there were 5795 BACs mapped to the genome using short BAC end sequences, and 6973 BACs mapped onto the genome using long BAC end sequence. Finally, the blast results generated by short and long BAC end sequences were integrated, and in total 8262 BACs were mapped to the genome with the rate of successfully mapped BACs of 89.65%.

Table 2 Anchoring the Broomcorn millet clones to the longmi 4 genome using BESs

Categories	Short BESs	Long BESs	Integrity
BAC end sequences			
Clones in superpool	9216	9216	
Clones with successful BESs	9126	7725	
with paired BESs	7790	3890	
with single-end BESs	1336	3835	
only forward	921	2189	
only reverse	415	1646	
Total successful BESs	17811 (96.63%)	12721 (66.17%)	
Alignment			
Aligned BESs with repeats unmasked	14862 (83.44%)	12120 (95.28%)	
Single hit BESs	12971	12120	
Multi-hit BESs	1891	0	
Aligned BESs with repeats masked	7760	10626	
Single hit BESs	7295	10626	
Multi-hit BESs	465	0	
Anchoring to reference sequences			
Clones anchored to single sites	5795	6973	8262
Clones anchored with single BES	2507	3907	2871
Clones anchored with paired BESs	3288	3066	5391

To verify the accuracy of BAC locations on the genome, 55 BAC clones were randomly picked from 384-plates for BAC end sequencing using Sanger method. The 55 paired end sequences after quality control were blasted with the above short and long BESs, and the genome. Finally, 54 BAC clones were consistent and only one BAC (32G16) end sequence was inconsistent. Checking the 32G16 BAC end sequences, we discovered that this clone had two long forward BESs, one short forward BES, and no reverse BES. Only one long BES and one short BES were perfectly identical to the Sanger end sequence. Another long BES confused the mapping, which can be removed by weighting. In summary, the accuracy of the BAC mapping approach in this study was extremely high.

The distribution of 5391 BACs mapped to the genome by paired BAC end sequences was counted (Table 3). These clones covered a total of 432.47 Mb of chromosomes with a total coverage of 50.97%. Among the 18 chromosome sequences of the cultivar longmi4 genome,

there were 829 gaps, of which 308 gaps were covered by our BACs. These BACs are valuable resources for further improvement of the genome.

Table 3 The location result of BAC on broomcorn millet each chromosome

Chr.	Length (Mb)	No. gaps	No. BAC	BACs Covered	
				length (Mb)	No. gaps
chr 1	69.18	85	439	35.25	34
chr 2	61.15	59	379	31.04	19
chr 3	57.97	50	394	31.03	20
chr 4	56.29	34	359	27.90	14
chr 5	54.13	52	361	29.86	19
chr 6	52.84	46	358	28.24	16
chr 7	51.23	67	286	23.67	27
chr 8	48.26	29	339	26.49	8
chr 9	45.11	70	240	20.96	21
chr10	44.65	53	308	25.44	28
chr11	43.18	30	267	21.90	14
chr12	42.47	30	254	20.79	11
chr13	40.72	50	261	21.11	17
chr14	38.49	32	269	20.23	10
chr15	34.36	34	213	18.14	8
chr16	33.61	45	212	16.72	15
chr17	32.99	25	226	18.26	10
chr18	32.24	38	199	15.32	17
unplaced	9.48	0	27	-	0
Total	848.47	829	5391	432.47	308

To more accurately assess the quality of broomcorn millet BAC library, including the average size of inserted fragments and contamination of organelle DNA, we carried out further analysis. The actual fragment sizes of these BACs presented the Gaussian distribution, with an average insert size of 123.48 kb, which is lower than prediction by pulsed-field electrophoresis mentioned at the section of BAC library constuction (Fig. 4).

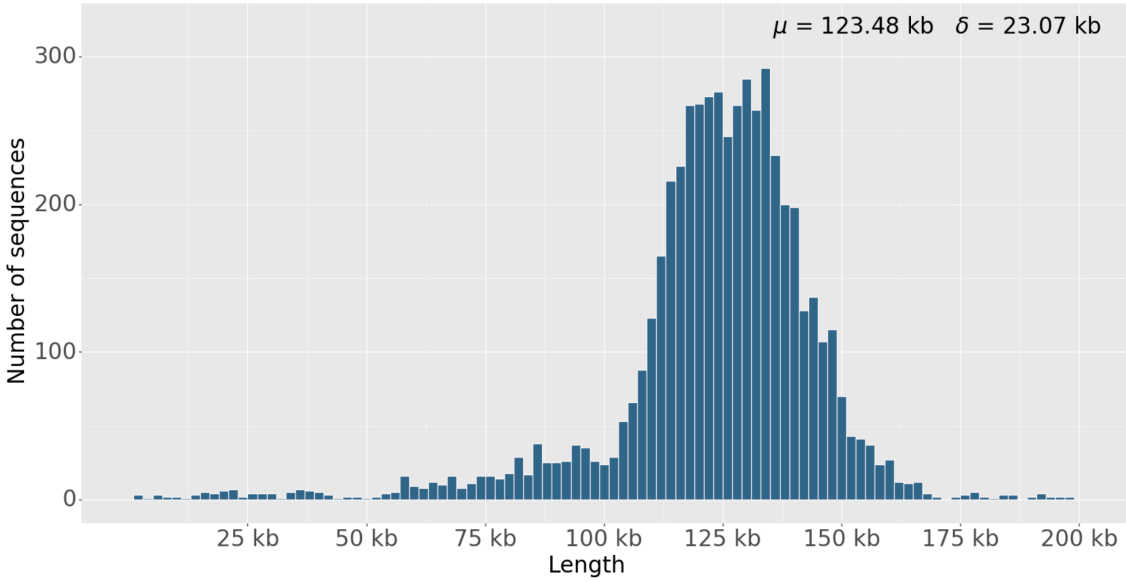

Fig. 4. the statistics of the located BAC lengths by paired BESs

2.5 Presentation of BAC location by JBrowse

In order to effectively use the BAC resource, quickly and easily retrieve BAC clones and view other annotaion information, we established a resource website and employed a lightweight JBrowse to display the broomcorn millet genome and BAC resource information (Fig. 5).

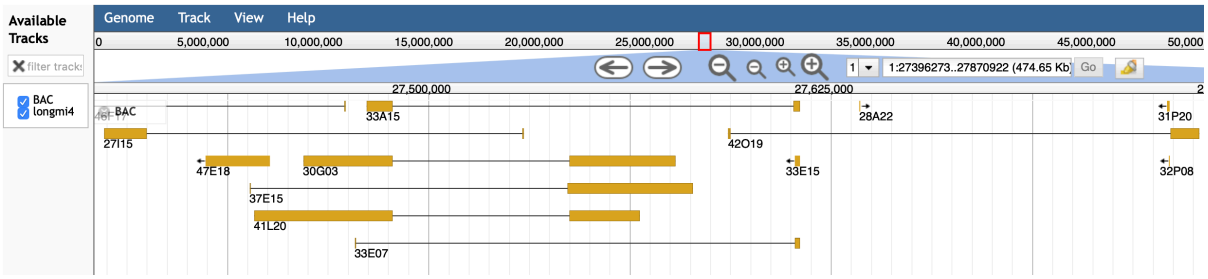

Fig. 5. The presentation of broomcorn millet BAC location in JBrowse. In BAC trace, two yellow rectangles (or vertical lines) which means BESs were connected by a horizontal line to indicate the position of the BAC with the paired BESs. one yellow rectangle (or vertical line) followed by a horizontal arrow indicate the BAC clone located by a single end, and the direction of arrow defer to the extension of the BAC.

2.6 Organelle genomes in the BAC library

To assess the proportion of organelle genomes in the BAC library, we aligned all BESs in this study to the chloroplast (CM009689) and mitochondria (grass) sequences by local blast. A total of 65 BAC clones were mapped to the chloroplast with more than 99% of similarity, of which 14 BAC clones were determined by paired BESs (25C07, 25N09, 27F16, 27O19, 29B02,

29K11, 32J09, 32N16, 33D23, 33K24, 39C22, 42C10, 46C16, 48C03). A total of 16 BAC clones had homologous BES sequences with mitochondrial genomes of Gramineae, of which 11 BAC end sequences were simultaneously aligned to the chloroplast and 3 BAC end sequences were simultaneously aligned to the nuclear genome. Only two BAC end sequences were aligned to neither the nuclear genome nor the chloroplast (36A05 and 48B15).

2.7 Extracting and *de novo* assembling unmapped reads

There were 3.2% of unmapped reads after mapping the clean reads to the longmi 4 genome. For investigate the information in these unmapped reads, Velvet was employed for assembling with kmer=121. We got 832 contigs which length were longer than 500bp and the GC content, cumulated length and N50 is 55.27%, 2.2Mb and 5579bp, respectively. These contigs were of high complextity evaluated by a online tool Prinseq (Fig.6).

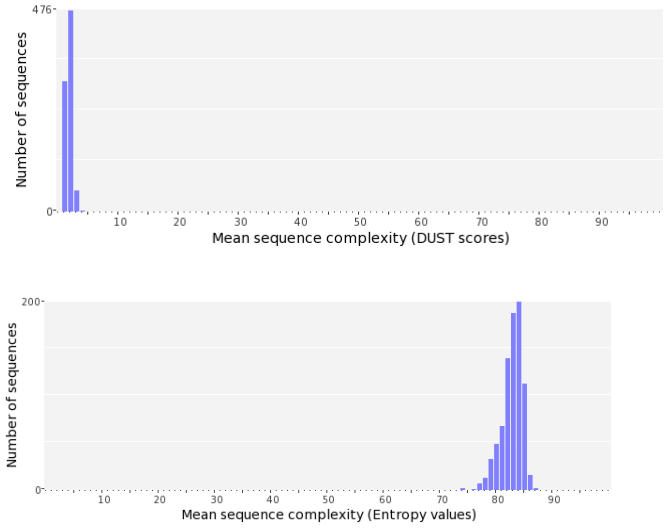

Fig. 6. DUST complexity scores and Entropy complexity scores of the contigs from unmapped reads. A online tool Prinseq was used to evaluate the complexity of the assembled unmapped contig (DUST <7 and Entropy>70). Low DUST scores (<7) and high Entropy scores (>70) indicate high complexity.

2.8 Homologous analysis of the unmapped contigs

We remained 301 homologous contigs after aligning the contigs above with NCBI nucleic acid database (NT) by Blastn. According to the taxid, there were 56 species, 18 of them were Gramineae and the rest were non-Gramineae. The 18 Gramineae species involved 137 contigs, 19 contigs were mitochondrial genomes, the rest were nuclear genomes and non chloroplast

genomes were found (Fig.7). According to the number of the contigs, the top five species of the Gramineae were Switchgrass, Millet, Rice, Maize and Sorghum.

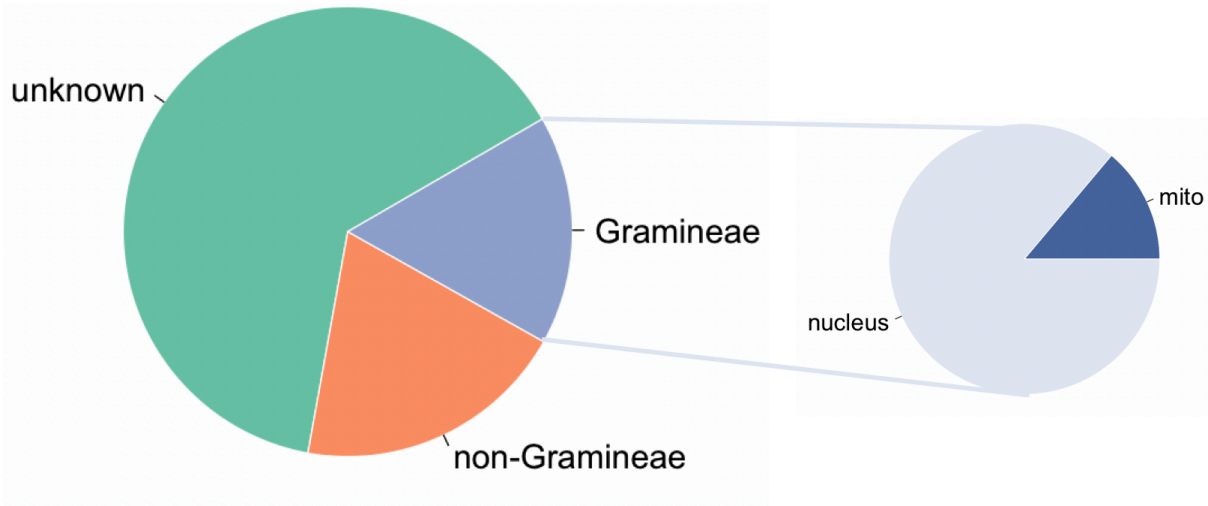

Fig. 7. The classification distribution of the unmapped broomcorn millet sequence contigs

Excluding the non-Gramineae contigs, the rest 668 unmapped contigs were aligned with longmi4 genome by Blastn. We get 659 homologus contigs, of which 343 were more than 50% coverage and 316 were less than 50% coverage , and the rest were not mapped.

2.9 Characterization of species-specific sequences

According to the results, 16.4% of the unmapped contigs may belonged to Gramineae. Among these contigs, 86.1% could map to the nuclear genome of Gramineae, suggesting that these contigs may be the species-specific sequence or the structural variation region. Candidate specific sequences were: (1) excluding contigs which were more than 50% of matching coverage with longmi4 genome; (2) excluding contigs which were mapped to mitochondria and chloroplast genome and non-gramineous plants. Finally, 104 contigs were obtained. The total length of 104 contigs was 193757 bp, the GC content was 46.39%, the N50 was 2595 bp, and the maximum length was 9886 bp.

By blasting 104 unmapped contigs with located BAC contigs parsing by CAPSS strategy, 64 unmapped contigs which were more than 95% of matching coverage were obtained. These 64 contigs were a total length of 118,214 bp and. The GC content was 46.02% and N50 was 2525 bp.

Augustus, an online soft was used to determine whether these contigs contain gene coding sequences. The Organism parameter use Oryza brachyantha and the rest parameters were default. Thirteen protein sequences were predicted in 10 contigs. When aligned with NR

database by blastp, we get 3 protein sequences and the sequences similarity ranged from 48.05% to 77.88%.

3 Discussion

We developed a high-efficient, low-cost BAC end sequence identification pipeline which can carry out ten thousand paired BAC end sequences at each time， to map clone location on the chromosomes and screen clone with species-specific sequences.

Bacterial artificial chromosome (BAC) library is a powerful resource for genome sequencing, gap filling, genome assembly assessment, target gene screening, functional genomics research, transformation of large DNA segments, recombineering, and DNA storages of the endangered species. Paired end sequences from BAC libraries are usally used to detect assembly error or assist assembley [17]. The most conventional and convenient approach to decode BAC end sequence profilles is Sanger sequencing method, but its one-by-one style is laborious, time-consuming and expensive. Fortunately, a few of efficient approaches had developped, based on next generation sequencing technologies, such as pBACcode, BAC-anchor[18, 19]. pBACode determined paired BESs by a pair of random barcodes flanking the cloning site. BAC-anchor identified paired end reads containing large internal gaps through specific restriction enzyme cutting sites, which BAC DNA were completely digested and subsequently self-ligated. We previously developed a high-throughput approach for long accurate BAC end sequence profile by PacBio sequencing technology [20]. However, these approaches focused on paired BES profiles, but it’s hard to trace BAC end suqence of specific clone in 384-plates well. Our approach applied the characteristics of row-cross-column in two dimensions array and cost-effective Illumina platform to generate BAC end sequences and assign to physical wells in 384-plates. These BAC end sequences on average are longer and accurater than those generated by Sanger method and others, owed to the assembly by SPAdes assembler[21]. After alignment, 89.65% of clones were successfully associated with the broomcorn millet longmi 4 genome. Those clones can be uesd to close the genome gaps to imporve the consistency.

In conventional whole genome sequencing project, sequencing coverage usallly is at least 30x, not more than 200x. Higher coverage will exist more equences that are generated by PCR mutations or sequencing errors and generates contigs with shorter N50. We added index into each secondary pool and mixed 96 secondary pool as a sequencing library that run in a Illumina

flow cell for greatly reducing the cost of sequencing. As a result, the average valid sequencing depths of row pools and column pools were 45x and 78x, respectively. We have to mention that the reads dated from *E. coli* genome should be excepted before estimating the coverages. *E. coli* genomic DNA was difficult to completely isolate with BAC DNA in the BAC plasmid preparation process. Therefore, in our experience, without special steps, there is about 40% reads form E. coli genome on average.

In process of the extraction of BAC end sequences, we designed two methods: a method called short BES pathway that BESs were shorter but more comprehensive, and another method called long BES pathway that BESs were longer. The BAC end sequences at the intersection were assigned to the 384-plate wells by the characteristics of the row-cross-column of the secondary pools. The BAC end sequence assignment at the intersection site is affected by many factors, such as the BAC clone overlapping rate and the correct rate of the sequences. The BAC clone overlapping rate in the superpool is the most important factor. When two same BAC clones appear in different rows and columns, we can filter them by our previous method [22]. However, if there are more than two same BAC clones that all of them contain in a row or column pool, our method will filter out all potential BESs, so that the BESs of the intersection well will absent. In the process of BES assignment, the flow of the forward and reverse BESs are completely independent. Because of the fault of assigned BESs, in the absence of a reference genome, no direct evidence can determine if the forward and reverse BESs from the BAC of the intersection well is a pair of BESs. If there was a high-qualty genome, it is easy to assess whether the two BESs derive from the same BAC by aligning loci. However, the variety of BAC library is inconsistent with the species in which the reference genome is obtained, the results of the alignment that do not satisfy the location requirement of a pair of BESs will be discarded. Wei et al. developed a method called pBACode, which adds a pair of barcodes on both sides of the BAC vector cloning site so that the BESs can be paired by paired barcode information [18]. Although a few of clone's BESs may be assigned improperly, the fault is random because of randomly clone picking. those clones with assigned BESs provided a valuable BAC resource for the improvement of the broomcorn millet genome and functional genomics research.

The plant organelle genome consists of the chloroplast and mitochondrial genomes. The chloroplast genome is generally around 150 kb, while the mitochondrial genome size varies widely, typically between 200 kb and 750 kb [23]. Plant cells contain the abundance of chloroplasts and mitochondria. During BAC library construction, Triton® X-100 was added to disrupt the chloroplast and mitochondrial structure, reducing organelle contamination in the

BAC library. In the study, the presence of a small number of chloroplast genomes was found by BESs, while the mitochondrial genome was almost absent, causing by sedimentation coefficient difference whose chloroplast is similar with the nucleus and the mitochondria is much lower than nucleus. The chloroplast genome of broomcorn millet has been reported, but the mitochondrial genome has not been reported yet [24]. Its mitochondrial genome can be decoded by isolating mitochondrial organelle for extracting DNA for sequencing and assembling and extracting homologous sequences with the mitochondria from the original data of whole genome sequencing [25].

The Illumina reads unmapped to longmi4 genome were explored for the hidden biological information and contamination, including deleted and species-specific sequences. Further analysis of unmapped reads can improve the quality of genome assembly and expand comprehension of species diversity. Blastn analysis of deleted or specific genomic sequences can provide a preliminary estimate of the origin of contig. However, this judgment is based on the results of homology comparison, and the authenticity of the results cannot be effectively judged. This seemingly correct result may be due to the incorrect assembly of unmapped reads.

By high-throughput sequencing and mapping in the secondary pool of BAC libraries, we can find the coordinates of genome deletion sequences in the broomcorn millet BAC libraries, which provides an effective method for searching genome deletion sequences or specific sequences in unmapped reads.

At the same time, if we find genes that play an important role in biology, we can quickly find BAC clones containing this gene, and we can quickly obtain experimental materials for further research and analysis. At the same time, because BAC contains a long DNA fragment (about 120 kb), it is also convenient for us to quickly analyze the upstream and/or downstream DNA elements of interested genes or adjacent genes.

4 Methods

4.1 Plant materials, growth conditions and BAC library construction

The seeds of broomcorn millet were provided by professor Chen and grown at dark conditions under 25°C. The young leaves of seedlings were harvested and mixed, frozen immediately in liquid nitrogen for the extraction of genomic DNA. High molecular weight genomic DNA was extracted and BAC library was constructed following the previous protocol. The main process as following: The young leaves of seedings were harvested and grinded in

liquid nitrogen for isolating cell nuclei. Cell nuclei were mixed with the equivalent of 1% Low melting temperature agarose to form plugs. Plugs were incubated with proteinase K solution for about 48 hours, washed with phenylmethylsulphonyl fluoride solution, finally stored in 4°C freezer with TE. The partial digestions of DNA plugs with dilution HindIII were completed. DNA fragments ranging from 100 kb to 200 kb were recovered from pulse field gel electrophoresis gel and ligated with pIndigoBAC536-S vectors. Ligated DNA was mixed and electroporated with electro-competent DH10B cells. The white colonies were picked up and stored in 384-plates.

4.2 Pool construction

24 384-plates were chosen form BAC library of broomcorn millet. These plates were arranged a 2-dimension superpool. The superpool contains 96 row pools and 96 column pools. A total of 192 pools were processed independently for high quality DNA extraction in AxyGen AxyPrep Easy-96 plasmid kit. Then, these BAC DNA were completely digested with ATP-Dependent DNase (Epicentre) for removing the DNA from *E. coli*. After digestion, these DNA were sheared in the Bioruptor to an average of 500 bp. During blunt-end repair, overhanging 5' and 3' ends were filled in or removed by T4 DNA polymerase. 5'-phosphates were attached using T4 polynucleotide kinase. Tail-A were added using Taq DNA polymerase. Then, adapters were ligated to both ends of the molecules using T4 DNA ligase. This ligation production was cleaned using MagBead DNA Purification Kit (Sangon Biotech, shanghai, CN). Sequencing adaptor were added using PCR. Finally, the production of KOD PCR were cleaned again.

4.3 Illumina sequencing and BAC end sequence analysis

Sequencing were performed via the Illumina HiSeq 2000 with 150-bp paired-end protocol (Genewiz, suzhou, CN). Trimmotatic was used to filter and trim raw reads. FastQC was used to assess data quality. After adapter filtering and quality assessment, BBMAP/demuxbyname script was employed for deconvolution depending on unique index sequence in each pool. A python script was used to extract target paired-end reads that covers BES-VES site. These target reads extracted would be assembled by Cap3 form consensus, called "short BES". SPAdes was used to directly assemble pool reads to form contigs. Then, contigs containing BAC vector sequence were extracted using python script, which is called "long BES". Using

Blastn to align BES from raw and column pools and assign them to the BAC at the intersection. False-positive BES were removed depending two methods.

4.4 Validation the correction of parsing BES

Analysis results of BAC end sequences were validated by Sanger sequencing. Fifty-five BAC clones were randomly selected and their DNA were extracted by an improved alkaline lysis protocol. Sanger sequencing was accomplished using BAC-F (5'-AACGACGGCCAGTGAATTG-3') and BAC-R (5'-GATAACAATTTCACACAGG-3') primers from pIndigoBAC536-S vector backbone. The BAC end sequence from Sanger sequencing align to BES from the analysis results of illumina data using Blastn.

4.5 Anchor BAC onto broomcorn millet longmi4 genome

The genome sequence of broomcorn millet longmi4 was downloaded from NCBI genome database GCA_002895445, which was submitted by researchers from China Agricultural University. All BESs were separately blasted (qcov_hsp_perc=99, perc_identity=99, outfmt=6, culling_limit=1) with the genome sequences. The output results were further converted to a GFF3 format file using a Python script. In this script, following conditions were setted. If both forward and reverse BESs in each clone were mapped to the same chromosome, and their orientations were opposite and their interval lengths were less than 250 kb, such clones were recorded in GFF3 format file with three lines. If either forward or reverse BES was uniquely mapped to chromosome, such clones also were recorded in GFF3 format file with two lines. The GFF3 file was sorted using GFF3sort and presenting with JBrowse in our website (http://gresource.hzau.edu.cn/resource/resource.html).

4.6 Unmapping reads analysis

The clean reads were mapped on broomcorn millet longmi4 genome using Bowtie2, and the unmapped reads further were extracted using samtools (the parameter is view -f 4). Velvet was employed for assembling these unmapped reads to form unmapped contigs (length was larger than 500 bp). The assembled unmapped contig was analyzed locally by Blastn. Firstly, compared with NCBI Nucleic Acid Database (NT), the best hits of each contig were sorted according to Bitscore and Pidnet, which the match length was more than 150 bp. The homologous species of contigs were classified according to taxid. Secondly, the unmapped

contigs with high similarity (Identity > 90 and length/qlen > 0.9) was excluded by comparing with the broomcorn millet longmi4 genome sequence. Finally, the Blastn results were counted, excluding unmapped contig, which was similar to the genome sequences of non-gramineae plants, organelle genomes and millet. The remaining unmapped contig was the specific or missing sequence of the candidate for broomcorn millet. The parameters used for Blastn alignment were： -outfmt "6 qseqid sseqid qlen length bitscore pident evalue qcovs staxids qstart qend sstart send stitle" -evalue 1e-20.

By comparing the screened unmapped contig with BAC contigs parsing through CAPSS strategy, the matching coverage of unmapped contig is more than 95%. Thus, the location of unmapped contig in the BAC library is determined. Sequence characteristic analysis, statistical length distribution, GC content distribution and repeat sequence analysis were carried out for the obtained specific or deleted sequence of broomcorn millet (Using online tools Tandem Repeats Finder on http://tandem.bu.edu/trf/trf.basic.submit.html and RepeatMasker on http://www.RepeatMasker.org/cgi-bin/WEBRepeatMasker). Online tools Augustus was employed to predicted gene (http://bioinf.uni-greifswald.de/augustus/submission.php).

5 Data availability

The source codes are available at https://github.com/xuweixw/broomcorn-millet-BAC-library. Illumina sequencing data are available at Sequence Read Archive (SRA) under the accession PRJNA576359.

References

1. Kalinova J, Moudry J. Content and Quality of Protein in Proso Millet (Panicum miliaceum L .) Varieties. Plant Foods Hum Nutr. 2006;61:45–9.

2. Washburn JD, Schnable JC, Davidse G, Pires JC. Phylogeny and photosynthesis of the grass tribe Paniceae. Am J Bot. 2015;102:1493–505.

3. Baltensperger DD. Progress with proso, pearl and other millets. 2002.

4. Shi J, Ma X, Zhang J, Zhou Y, Liu M, Huang L, et al. Chromosome conformation capture resolved near complete genome assembly of broomcorn millet. Nat Commun. 2019;10:1–9. doi:10.1038/s41467-018-07876-6.

5. Zou C, Li L, Miki D, Li D, Tang Q, Xiao L, et al. The genome of broomcorn millet. Nat Commun. 2019;10:1–12. doi:10.1038/s41467-019-08409-5.

6. He X, Cicek AE, Wang Y, Schulz MH, Le HS, Bar-Joseph Z. De novo ChIP-seq analysis. Genome Biol. 2015.

7. Isakov O, Modai S, Shomron N. Pathogen detection using short-RNA deep sequencing subtraction and assembly. Bioinformatics. 2011.

8. Hajirasouliha I, Hormozdiari F, Alkan C, Kidd JM, Birol I, Eichler EE, et al. Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. Bioinformatics. 2010.

9. Kavak P, Lin YY, Numanagić I, Asghari H, Güngör T, Alkan C, et al. Discovery and genotyping of novel sequence insertions in many sequenced individuals. In: Bioinformatics. 2017.

10. Chen R, Lau YL, Zhang Y, Yang W. SRinversion: A tool for detecting short inversions by splitting and re-aligning poorly mapped and unmapped sequencing reads. Bioinformatics. 2016.

11. Daniels PW, Mukherjee A, Goldman AS, Hu B. A set of novel CRISPR-based integrative vectors for Saccharomyces cerevisiae. Wellcome Open Res. 2018;3:72.

12. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics. 2009.

13. Saxena RK, Edwards D, Varshney RK. Structural variations in plant genomes. Briefings Funct Genomics Proteomics. 2014.

14. Tattini L, D'Aurizio R, Magi A. Detection of genomic structural variants from next-generation sequencing data. Frontiers in Bioengineering and Biotechnology. 2015.

15. Lee T, Kim MY, Ha J, Lee SH, Scofield D. Detection of large sequence insertions by a hybrid approach that combine de novo assembly and resequencing of medium-coverage genome sequences. Genome. 2018.

16. Fan X, Abbott TE, Larson D, Chen K. BreakDancer: Identification of genomic structural variation from paired-end read mapping. Curr Protoc Bioinforma. 2014.

17. Deng Y, Pan Y, Luo M. Detection and correction of assembly errors of rice Nipponbare reference sequence. Plant Biol. 2014;16:643–50.

18. Wei X, Xu Z, Wang G, Hou J, Ma X, Liu H, et al. PBACode: A random-barcode-based high-throughput approach for BAC paired-end sequencing and physical clone mapping. Nucleic Acids Res. 2017;45:1–17.

19. Yang X, Yang Y, Ling J, Guan J, Guo X, Dong D, et al. A high-throughput BAC end analysis protocol (BAC -anchor) for profiling genome assembly and physical mapping . Plant Biotechnol J. 2019;:1–9.

20. Zhaozhao D, Tong L, Jiadong L, Zhifei H. High-throughput long paired-end sequencing of a Fosmid library by Pacbio. Plant Methods. 2019.

21. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 2012;19:455–77.

22. Pan Y, Wang X, Liu L, Wang H, Luo M. Whole Genome Mapping with Feature Sets from High-throughput Sequencing Data. PLoS One. 2016;11:1–17.

23. Gualberto JM, Newton KJ. Plant Mitochondrial Genomes: Dynamics and Mechanisms of Mutation. Annu Rev Plant Biol. 2017;68:225–52. doi:10.1146/annurev-arplant-043015-112232.

24. Cao X, Wang J, Wang H, Liu S, Chen L, Tian X, et al. The complete chloroplast genome of Panicum miliaceum. Mitochondrial DNA Part B Resour. 2017;2:43–5.

25. Tian X. The Rice Mitochondrial Genomes and Their Variations. PLANT Physiol. 2005;140:401–10.

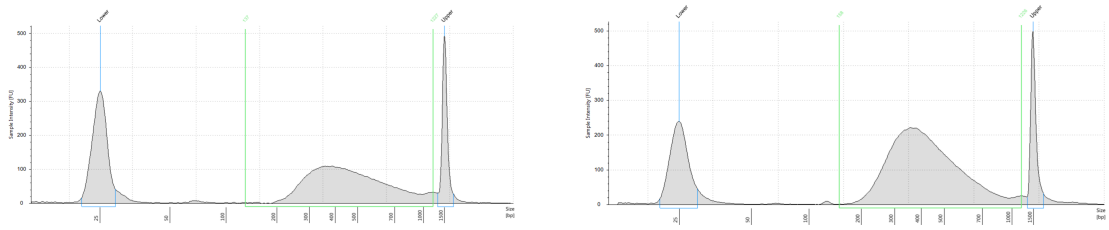

Fig. S1. the distribution of DNA fragment length in library X and Y. The peak in the middle of each figure is the length range of DNA library. The peaks on the left and right are 25 bp and 1500 bp DNA standard respectively. (A) Sequencing library X in the left, (B) Sequencing library Y in the right.

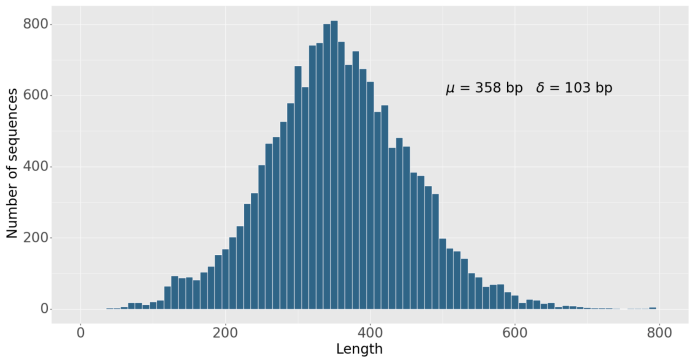

Fig. S2. the distribution of short BAC end sequence lengths

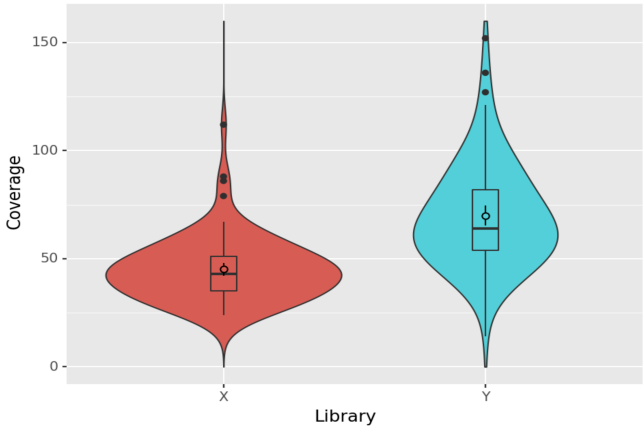

Fig. S3 the distribution of valid coverage in each pool. The average length of BAC inserted fragment is 132 kb (estimated by PFGE) and the length of BAC vector is 7 kb. The coverage of BACs in each pool is calculated with 139 kb.

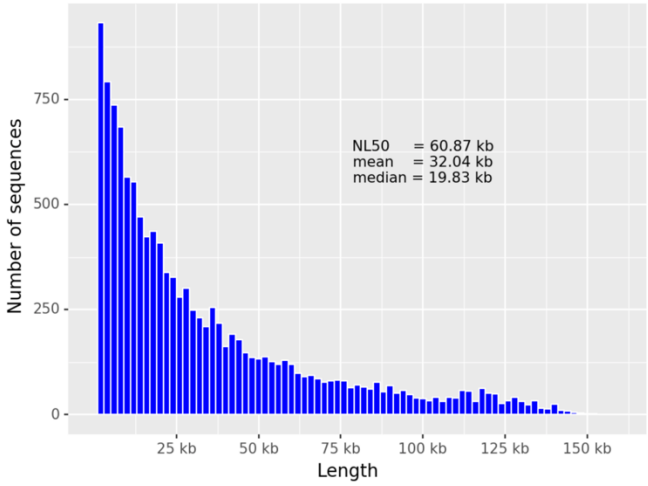

Fig. S4. the distribution of long BAC end sequence lengths

Table S1 the genome information of two broomcorn millet varieties

categories	Longmi4	landrace
No. NCBI BioProject	PRJNA429322	PRJNA431363
No. NCBI Assembly	GCA_002895445.2	GCA_003046395.1
Tech. type	Illumina、PacBio、HiC、BioNano	Illumina、PacBio、HiC、binmap
Total contigs	1308 (838.9 Mb)	5,541 (838 Mb)
N50 contigs	2.55 Mb	396 kb
Total scaffolds	475 (848.4 Mb)	1309 (855 Mb)
N50 scaffolds	48.26 Mb	46.66 Mb
GC content	46.8%	46.8%
Numbers of genes	63,671	65,573
Publication	(Shi et al 2019)	(Zou et al 2019)

Table S2 the result of short BESs location

	0	1	2	3	4
Forward BES	502	8183	487	89	2
Reverse BES	1011	7897	302	6	0

Table S3 the result of long BESs location

	0	1	2	3	4
Forward BES	3138	5454	591	31	2
Reverse BES	3680	5108	410	17	1