Journal of Genetics (2022)101:39
https://doi.org/10.1007/s12041-022-01382-x

RESEARCH ARTICLE

Genome survey and development of 13 SSR markers in *Eucalyptus cloeziana* by NGS

XIN-YUAN LIANG[1], TIAN-DAO BAI[1], JIAN-ZHONG WANG[2] and WEI-XIN JIANG[1]*

[1]*College of Forestry, Guangxi University, Nanning 530000, People's Republic of China*
[2]*Dongmen Forest Farm of Guangxi Zhuang Autonomous Region, Chongzuo 532108, People's Republic of China*
*For correspondence. E-mail: jwx_1985@163.com.

Received 23 December 2021; revised 2 July 2022; accepted 7 July 2022

Abstract. *Eucalyptus cloeziana* is a valuable timber tree species for its high durability and excellent sawmilling qualities. However, there is lack of complete genomic information for this plant, which severely constrains its genetic improvement. This study aim to survey the genome of *E. cloeziana* and determine the large-scale sequencing scheme of this species. Next-generation sequencing based on Illumina Hi-Seq X Ten platform was used to survey the *E. cloeziana* genome and its SSR markers development. We estimated the genome size to be 491.91 Mb and the heterozygosity rate to be 1.23%, with repetitive sequences accounting for 40.74%. The clean reads of *E. cloeziana* were assembled into 995,093 scaffolds (556,992,952 bp) with a N50 value of 2297 bp. In the GO database, the 10,172 genes annotated were matched to 50 functional gene groups in three categories of cell component, biological function and molecular function, respectively. Through KEGG pathways analysis, 10,802 genes were successfully annotated and 133 metabolic pathways were detected with the most abundant metabolism-related genes. Meanwhile, a total of 58,832 simple-sequence repeat (SSR) loci were identified in the *E. cloeziana* genome, and among them, dinucleotide repeats were the most abundant class. AG/CT, AT/AT, AAG/CTT were the three most frequent primitive types. Of the 50 genomic SSR primer pairs randomly selected for screening test, 13 showed polymorphism (PIC = 0.625). Three to nine alleles per locus (mean = 6.23) were observed, with the observed and expected heterozygosity at 0.317–1.000 and 0.276–0.838 across all the 44 *E. cloeziana* germplasms, respectively. Here, we report the genome information of *E. cloeziana*, and the novel 13 genomic SSR markers that were developed can be used as powerful tools for evaluating genetic diversity and population structure, and thus contribute to molecular breeding studies of *E. cloeziana* and other eucalypts.

Keywords. genome sequencing; genomic-SSRs; polymorphism; *Eucalyptus cloeziana*.

Introduction

Eucalyptus cloeziana F. Muell (Myrtaceae) is a timber tree species endemic to central and northern Queensland, Australia, with a range of 15°45′ ∼ 26°41′ S in latitude and 144°44′ ∼ 152°52′ E in longitude (FAO 1979). In addition to the advantages of common eucalypts such as fast growth, high yield and short rotation cutting period, more specially, its timber presents yellow–brown, heavy, strong and very durable heartwood (FAO 1979) and has thus been cultivated with the major aim to produce high-value solid wood products (Dickinson *et al*. 2000). *E. cloeziana* was first introduced to southern China in 1972, and further studies have been carried out mainly including the introduction test, provenance selection, pedigree test, correlation analysis between timber and growth properties (Huang *et al*. 2018). It is reported that the volume increment of superior *E. cloeziana* provenances at 9.5 years of age can reach up to 214.05 $m^3 \cdot ha^{-1}$ at State-owned Dongmen Forest Farm of Guangxi province, and this research was funded by a cooperative project between China and Australia from 1982 to 1989 (Qi 2002). The average basic density of 17-year old *E. cloeziana* was 0.706 $g \cdot cm^3$ (Li *et al*. 2012) and the heritability of a 25-year old individual was 0.136–0.342 (Wang *et al*. 2016). Early study indicated that significant variations were identified among open-pollinated progenies in breast-high diameter at different ages (Marques *et al*. 1996). High genetic diversity was found for a breeding population of *E. cloeziana* (Lv *et al*. 2020) and moderate differentiation was also observed between the populations from several distribution regions (Deng *et al*. 2019). For its excellent performance and great breeding value, it has been identified as a

Published online: 02 September 2022

precious timber tree in China (Li *et al*. 2012; Huang *et al*. 2018).

Genome survey sequencing through the next-generation high-throughput sequencing (NGS) technique is an important and cost-effective strategy for discovering extensive genetic and genomic information relating to the varied characteristics of organisms (Morozova and Marra 2008), and developing molecular markers with good accuracy for plant breeding (Ray and Satya 2014). With NGS, genomics research has simultaneously gained speed, magnitude and scope. However, complete genomic information on *E. cloeziana* has remained largely unknown. Thus far, most genomic researches for *Eucalyptus* are mainly focussed on very few commercially important species, such as *E. grandis* and *E. camaldulensis* (Hirakawa *et al*. 2011; Paiva *et al*. 2011; Myburg *et al*. 2014). Over the last decade, many articles about the development of SSR markers of *Eucalyptus* have been published worldwide (Schmid and Harper 1985; Pérez-Vega *et al*. 2010; Xu *et al*. 2019). In contrast, comprehensive studies about developing genomic or EST SSRs for *E. cloeziana* have rarely been published. Although the whole-genome analysis on *E. grandis* has been reported (Myburg *et al*. 2014), a previous phylogeny reconstruction in *Eucalyptus* (Steane *et al*. 2011) showed that *E. cloeziana* belongs to an independent subgenus far from other eucalypts. The lack of a reference genome of *E. cloeziana* severely constrains its enhancement in molecular biology. Meanwhile, SSR primers suitable for genetic analysis of *E. cloeziana* are still limited, which is difficult to meet the research needs of genetic diversity analysis and genetic map construction of *E. cloeziana*. In the present study, we conducted a genome survey of *E. cloeziana* based on Illumina Hi-Seq technology combined with K-mer analysis and genome annotation, and described the development and validation of a set of 13 polymorphic SSR markers. This study could be beneficial to accelerate the progress of genetic improvement and better utilization of *E. cloeziana* gene resources.

Materials and methods

Plant materials

Fresh leaf samples from a 7-year-old *E. cloeziana* were collected from the Guangxi State-owned Dongmen Forest Farm, located in Fusui County, Guangxi, China (22°17′ 22.30″N, 107°14′108.00″E). After cleaning and disinfecting the leaves with 70% alcohol, they were immediately stored in liquid nitrogen. Total genomic DNA was isolated by using a DNA extraction kit (Sangon Biotech, Shanghai, China) and sent to OE Biotech (Shanghai OE Biotech, Shanghai, China) in dry ice. Fresh healthy leaves of 44 adult trees were collected from one *E. cloeziana* progeny trial constructed by the Guangxi State-owned Dongmen Forest Farm, located in Fusui County, Guangxi Province, China (22°17′22.30″N, 107°14′ 108.00″E). The family seedlots of one trial (n = 44) were collected from natural mother stands of the Australian Tree Seed Centre (ATSC, Canberra, Australia).

Genome sequencing

The *E. cloeziana* DNA samples of suitable quality were randomly sheared into 350 bp fragments using an Ultrasonicator (Covaris, USA). Electrophoresis was used to recover the DNA fragments of required lengths before end-repair, followed by poly A-tail and sequencing adapters were added. The obtained fragments were used to construct two paired-end Illumina libraries and then sequenced with a read length of 2 × 150 bp using the Illumina Hi-Seq X Ten platform (Illumina, San Diego, USA). Clean reads were acquired for all subsequent information analysis, including the size of the genome, repetitive sequences, heterozygosity, etc. Before that, 250,000 pairs of reads selected randomly were used to search against the NCBI database using BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi). The species matching the largest number of reads were examined to investigate the potential contaminating effect.

Genome size estimation and genome survey

Every K-mer of each sequence reads was counted, and then the frequency of every K-mer was counted. K-mer frequency (depth of coverage) distribution follows the Poisson distribution (Chor *et al*. 2009; Nowak 2015). From the K-mer depth distribution, we obtained the peak value of depth, which is the average value and variance of the related Poisson distribution. A routine 17-mer frequency distribution analysis https://www.nature.com/articles/s41597-020-0441-7-ref-CR13 was performed according to the following formula: genome size = K-mer num/peak depth relatively (Marcais and Kingsford 2011). The existence of heterozygote and repeating sequences in genome affects the K-mer depth distribution. Therefore, the heterozygous frequency and repeat sequence can be roughly determined by the depth distribution of K-mer. While we could also deduce the heterozygous ratio and repeating sequences in genome based on K-mer analysis (Marcais and Kingsford 2011). The software SOAPdenovo (Xie *et al*. 2014) was used for the assembly of the refined paired-end sequencing reads. K-mer size = 41 was chosen for assembly with default parameters in SOAPdenovo to construct the de Bruijn graph. The guanine plus cytosine (GC) average sequencing depth was calculated by the 10-kb nonoverlapping sliding windows along the assembled sequence. The quality of the completeness of the assembly was assessed using the BUSCO v3.1.0 (Waterhouse *et al*. 2018) method based on a benchmark of 2120 conserved plant genes. Gene ontology (GO) functional and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses of putative genes were

performed via GMAP software (Wu and Watanabe 2005). Matches with an e-value < 1e−5 and > 40% sequence identity were selected.

Development of SSR markers

MIcroSAtellite identification tool (MISA) (http://pgrc.ipk-gatersleben.de/misa/) was adopted to detect SSRs in the assembled genome sequences. The parameters were set based on the minimum number of iterations: 6, 5, 5, 5 and 5 repeat units for di-, tri-, tetra-, penta- and hexa-nucleotides respectively. A compound SSRs was defined when the distance between the two SSR segments is less than 100 bp. The primer pairs were designed by using the Primer 3 software as follows: a final product size range of 100–350 bp, primer size of 18–22 bp, a primer GC content of 40–60%, and a primer annealing temperature of 58–60°C. A total of 50 primer pairs with motif length of more than 20 bp were randomly selected and synthesized. The PCR amplification was performed by using six different DNA templates of *E. cloeziana* for prescreening, while the polymorphism analysis was studied by 44 *E. cloeziana* germplasms. The PCR procedure was carried out in 10 μL volume containing 20 ng of genomic DNA, 0.25 μM of forward and reverse primers and 5 μL of 2× PCR Mixture (Takara, Beijing, China) with conditions as follows: denaturation for 6 min at 95°C followed by 35 cycles of 30 s at 95°C, 30 s for annealing at 58°C and 1 min at 72°C, and a final extension at 72°C for 5 min. The amplified products were resolved by 8% polyacrylamide gel electrophoresis (PAGE) and the bands were developed with $AgNO_3$ solution. Bands were detected and genotyped by using GelJ software (Heras *et al.* 2015) and then manually adjusted for analysis.

Data analysis

The number of alleles (*Na*), observed heterozygosity (H_O), expected heterozygosity (*He*), deviations from Hardy–Weinberg equilibrium (HWE), and linkage disequilibrium (LD) between loci were estimated by using GenAlEx v. 6.5 (Peakall and Smouse 2012). Polymorphism information content (PIC) of polymorphic loci was calculated using Cervus 3.0.7 software (Butrinowski *et al.* 2013).

Results

Sequencing and data cleaning of E. cloeziana

Qualified DNA were first randomly interrupted with a fragment length of 350 bp, and after removing adapters and primers ~75.82 Gb of raw data were obtained. After strict quality control, we obtained ~74.69 Gb of clean reads. The Q30 value was 89.96% and the proportion of valid bases is 98.50%. Figure 1 shows the proportion of single bases, which is usually used to detect if AT and GC are present separately. It can be seen that the content of A, G, C and T are close, indicating that the sequencing quality was good. We randomly selected 500,000 clean reads as a query sequence with basic local alignment search tool (BLAST) against the NT (Nucleotide Sequence Database) from the NCBI, and the species matching the largest number of reads were examined to investigate the potential contaminating effect. The top five species for comparison are *E. grandis* (137,772, 84.25%), and *Vitis vinifera* (5,832, 3.57%), *Cajanus cajan* (4,523, 2.77%), *Corchorus capsularis* (3,730, 2.28%) and *Cephalotus follicularis* (3,216, 1.97%). The BLAST results showed that the library was established without contamination from other species.

Genome size estimation by K-mer analysis

All the clean data were performed on K-mer analysis. A total of 66,687,003,174 different 17-mers were analysed and the peak of the depth distribution was at 133× according to the 17-mer frequency distribution (figure 2). The estimated genome size of *E. cloeziana* was 501.41 Mb, which was calculated via the formula *Kmer-num/Kmer-depth*. Then, the genome size was reformulated by excluding the K-mer error, and the revised genome size was 491.91 Mb. The gene heterozygosity rate was calculated to be 1.23% according to the K-mer curve distribution. Through calculating the percentage of 1.8-times the number of K-mers after the main peak over the total number of K-mers, we obtained the repetitive sequences accounted for 40.74% of the whole genome.

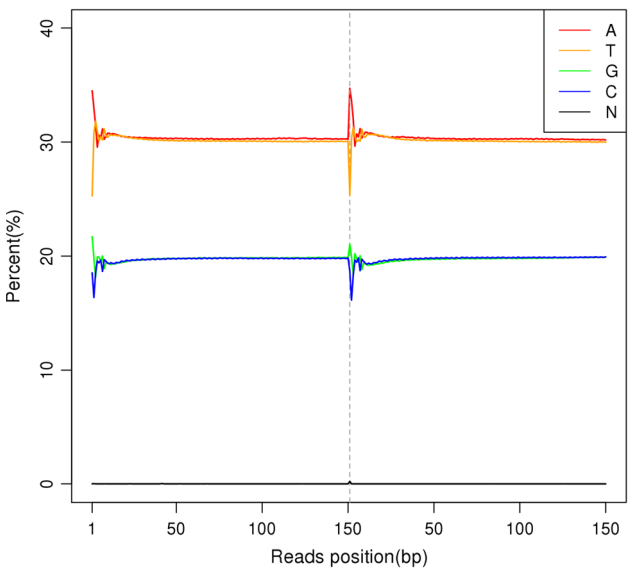

Figure 1. Base distribution of *E. cloeziana*.

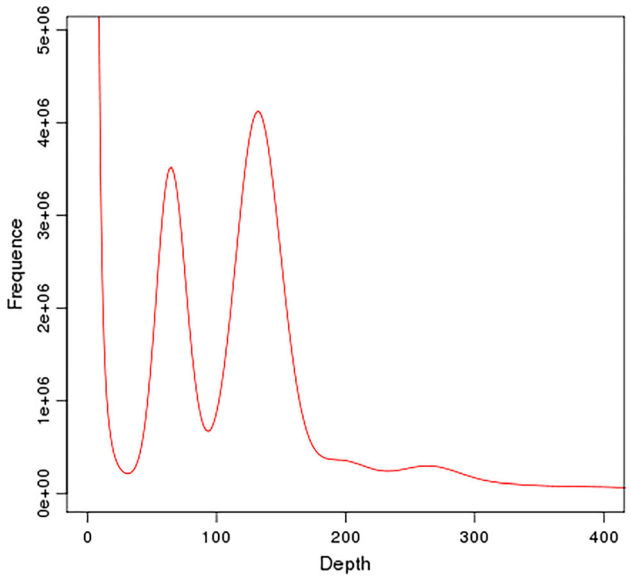

Figure 2. Distribution of 17-mer frequency of *E. cloeziana*. The *x*-coordinate is the depth of the K-mer 17, and the *y*-coordinate is the frequency of the K-mer species.

Genome assembly and GC content analysis

After filtering, a K-mer value of 41 was selected to construct the contig and scaffold. Through strict sequence alignment, the insert size of each library was confirmed. Scaffolds larger than 1000 bp were selected to avoid low-quality sequences and then 1,196,365 contigs (totaling 535.21 Mb) with an N50 of 1197 bp were generated. The longest contig was 93,276 bp. With the help of SSPACE, the genome assembly consisted of 995,093 scaffolds, with a total length of 556,992,952 bp. The N50 scaffold was 2297 bp and the longest scaffold was 212,254 bp (table 1). The average sequencing depth and GC content of the *E. cloeziana* genome were plotted along the assembled contigs, the length of which is more than 200 bp (figure 3). As shown in figure 3, most of the contigs were concentrated in the 25–55% GC content, from 50× to 120× average depth, and the average GC content was 39.63%. In this area, contigs had two gravity centres, which were located at an average depth of ~50× and 100×, respectively. Combined with the K-mer analysis, we inferred that the two gravity centres correspond to the heterozygosity and main peak, respectively. It can be seen from the figure that GC content distribution and contig coverage depth are relatively concentrated. There was no obvious abnormal dot distribution area in the figure. The BUSCO results (figure 4) showed that only 26.1% of the region had complete and single-copy genes (including 0.5% duplicated ones), 27.4% were partially matched BUSCO library and 46.5% were missing.

Table 1. Statistics of assembled *E. cloeziana* genome sequences.

Item	Contig (len)	Scaffold (len)	Contig (num)	Scaffold (num)
Total	535,208,608	556,992,952	1,196,365	995,093
Max length	93,276	212,254	–	–
Number ≥ 2000	–	–	41,873	45,258
N50	1197	2297	76,694	39,395
N60	726	1231	134,730	73,030
N70	424	639	231,623	136,595
N80	227	303	404,424	265,861
N90	145	150	725,671	553,540

Genome annotation

***GO function class*:** A total of 10,172 putative genes were idetified by the GO slim analysis and further classified into the categories of cellular component, biological processes and molecular function (figure 5). Specifically, 43.72%, 40.93%, and 15.35% of the genes were grouped under cellular component, biological process and molecular function, respectively. For cellular component, the two most represented categories were cell (8270) and cell part (8253), followed by genes involved in organelle (6646). For biological process, the most represented category was cellular process (6749), the second was genes involved in biological process (5662), followed by response to stimulus (2950) and biological regulation (2,599). For molecular function, the most two representative genes were those related to binding (5899) and catalytic activity (5195).

***KEGG metabolic pathway*:** There were 10,802 putative genes assigned to 133 KEGG pathways and divided into five categories, including metabolism, genetic information processing, environment information processing, cellular processes, and organismal systems and their 18 sub-categories (figure 6). For the metabolism, the genes involved in carbohydrate metabolism (1807, 16.73%) accounted for the largest proportion followed by lipid metabolism (1139, 10.54%) and amino acid metabolism (1017, 9.41%). There were four metabolic pathways related to genetic information processing, among which the genes involved in translation (1162, 10.76%) accounted for the largest proportion followed by folding, sorting and degradation (1076, 9.96%); The other pathways mainly involved signal transduction (831, 7.69%), transport and catabolism (785, 7.27%) and environmental adaptation (677, 6.27%).

Charactersitcs of genomic SSRs

After filtering the SSR sequences from the scaffold sequences (995,093) at both sides (less than 100 bp), a total of 58,832 SSRs were identified in the study (table 2). The frequency of SSR was 5.91% and the average distance of genomic SSR was 9.47 kb. Dinucleotide repeats (36,036) were the most abundant motifs, accounting for 61.25% of all SSRs except mononucleotides, followed by tri- (12,997,

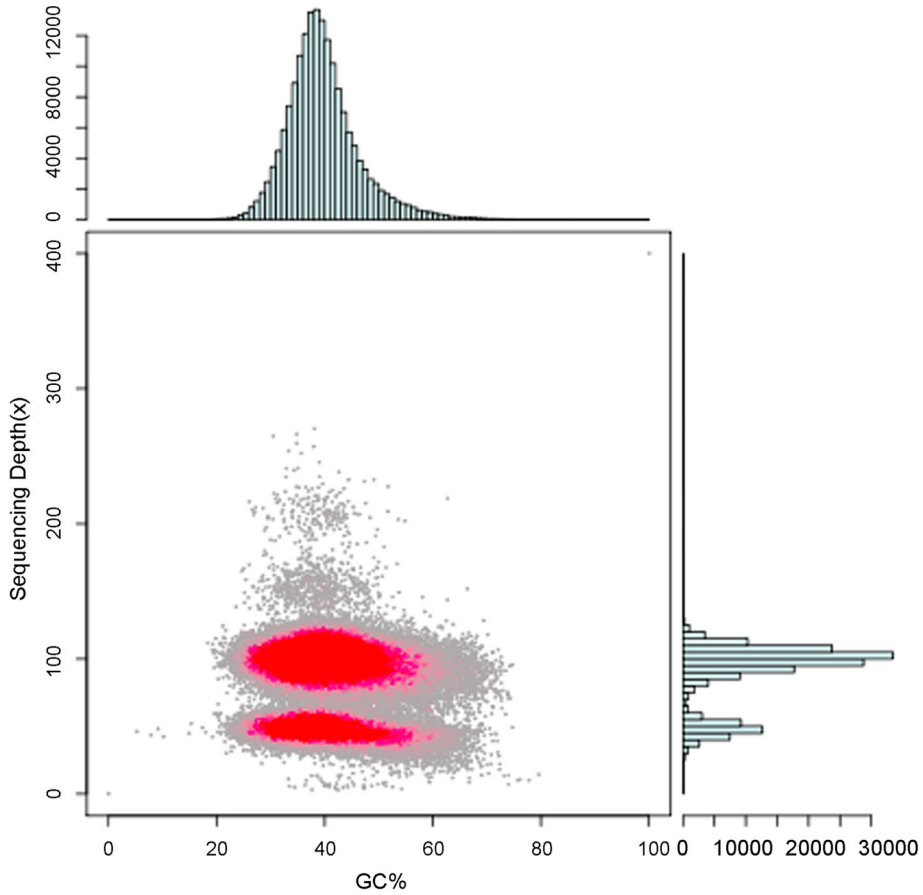

Figure 3. GC depth distribution of *E. cloeziana* genome contig. The *x*-axis is the GC content and the *y*-axis is the average sequencing depth. Each dot represents a contig; the GC content distribution is above; colour from red to gray indicates dot density from high to low.

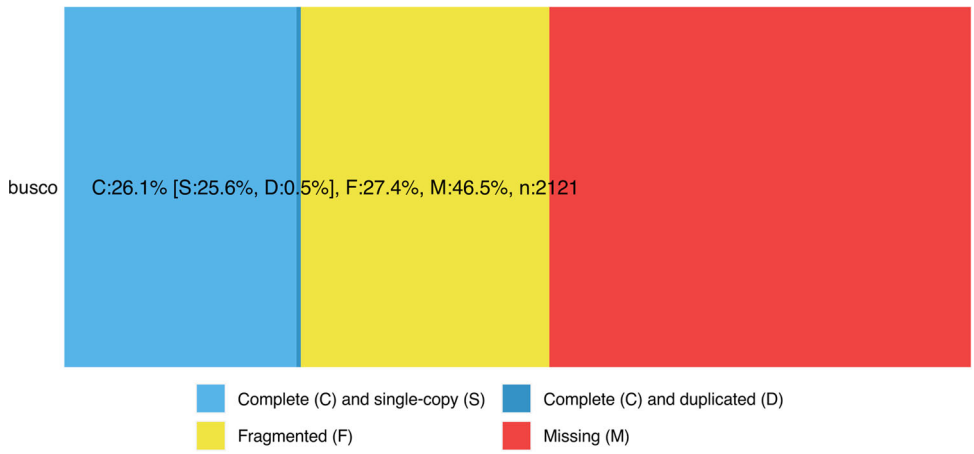

Figure 4. BUSCO quality assessments for *E. cloeziana* genome.

22.09%), tetra- (2,419, 4.11%), penta- (780, 1.33%) and hexanucleotide (380, 0.65%) repeats. Among the dinucleotide repeats, the predominant motif was AG/CT, accounting for 45.64%, followed by AT/AT repeats (11.95%) (table 2). Among the trinucleotides, the primary repeat was AAG/CTT, accounting for 7.49%, followed by AAT/ATT (4.41%) and AGG/CCT (2.6%). The repetitions of *E. cloeziana* SSRs were mainly from 5 to 11 times, accounting for 80% of the total number of SSR. For the most abundant dinucleotide repeats, the proportion of repetitions ≥ 12 times was 19.29%, including 985 SSRs with high repetitions (≥ 20 times).

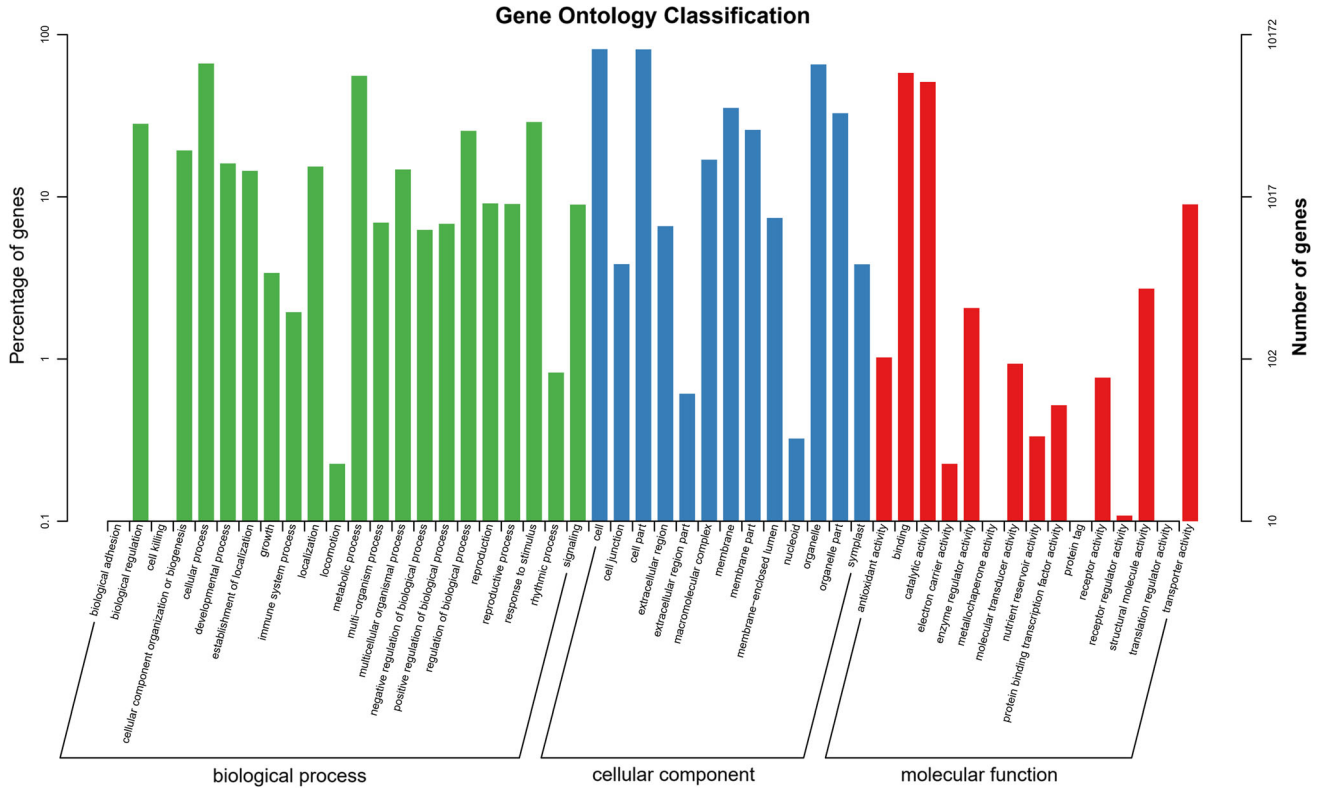

Figure 5. GO classification of predicted genes in *E. cloeziana*. The *y*-axis indicates the number of predicted genes in the GO databases. The *x*-axis indicates GO classification of *E. cloeziana* putative genes. Genes were assigned to three categories: cellular components, molecular functions, and biological processes.

Genomic SSR development

A total of 18,150 low-repeats (di- and tri-nucleotide) SSR sequences with fragment length $\geq$ 20 bp that met the primer design requirements were detected. To validate these SSR markers, 50 primer pairs randomly selected were synthesized and amplified for prescreening, and finally 13 novel SSR loci (table 3) showed a useful degree of polymorphism across 44 germplasms of *E. cloeziana*. The SSR summary statistics showed that the allele size ranged from 118 to 316 bp. There were from three to nine alleles across all loci, with an average of 6.23 alleles. The observed heterozygosity (*Ho*) and the expected heterozygosity (*He*) ranged from 0.317 to 1.000 (mean = 0.699) and from 0.276 to 0.838 (mean = 0.669), respectively (table 3). The PIC value ranged from 0.254 to 0.817 (mean = 0.625). Of the 13 genomic SSRs, 11 (84.62%) were demonstrated to have high polymorphism in 44 individuals (PIC $>$ 0.5). A total of five loci significantly ($P < 0.001$) deviated from HWE in these germplasms and we found no consistent deviation from linkage disequilibrium for any loci within the population ($P < 0.001$). Further, the corresponding sequences of these 13 SSRs were BLASTed against the GenBank nonredundant database using BLASTX (Altschul *et al.* 1997) (table 3).

Discussion

Genome survey gives a preliminary understanding of the genomic characteristics before carrying out the large-scale genome sequencing for one species. Genome size, usually described using the genomic content or DNA C-value, refers to the DNA content of the gamete genome. Based on the genome survey sequencing data combined with K-mer analysis, we could estimate the genome size of a nonmodel plant species, which is the basis for comparative and evolutionary genomics (Leitch and Leitch 2013). We can also comparatively analyse the genome size of different species and detect, identify, and grasp the regularity of the genome variation (Zhou *et al.* 2018). The genome size of *E. cloeziana* (491.91 Mb) is close to *E. pauciflora* genome (500 Mb) (Wang *et al.* 2020), larger than the *E. citriodora* genome (370 Mb) and *E. torelliana* (390 Mb) (Grattapaglia and Bradshaw 1994) but smaller than *E. grandis* genome (691.43 Mb) (Myburg *et al.* 2014), *E. camaldulensis* (654.92 Mb) (Hirakawa *et al.* 2011) and *E. urophylla* (650 Mb) (Grattapaglia and Bradshaw 1994). According to the new classification in *Eucalyptus* (Brooker 2000), the last three eucalypts with larger genomes belong to the subgenus *Symphyomyrtus*, the first two with lower genomes size belong to *Corymbia* while *E. cloeziana* is classified to the

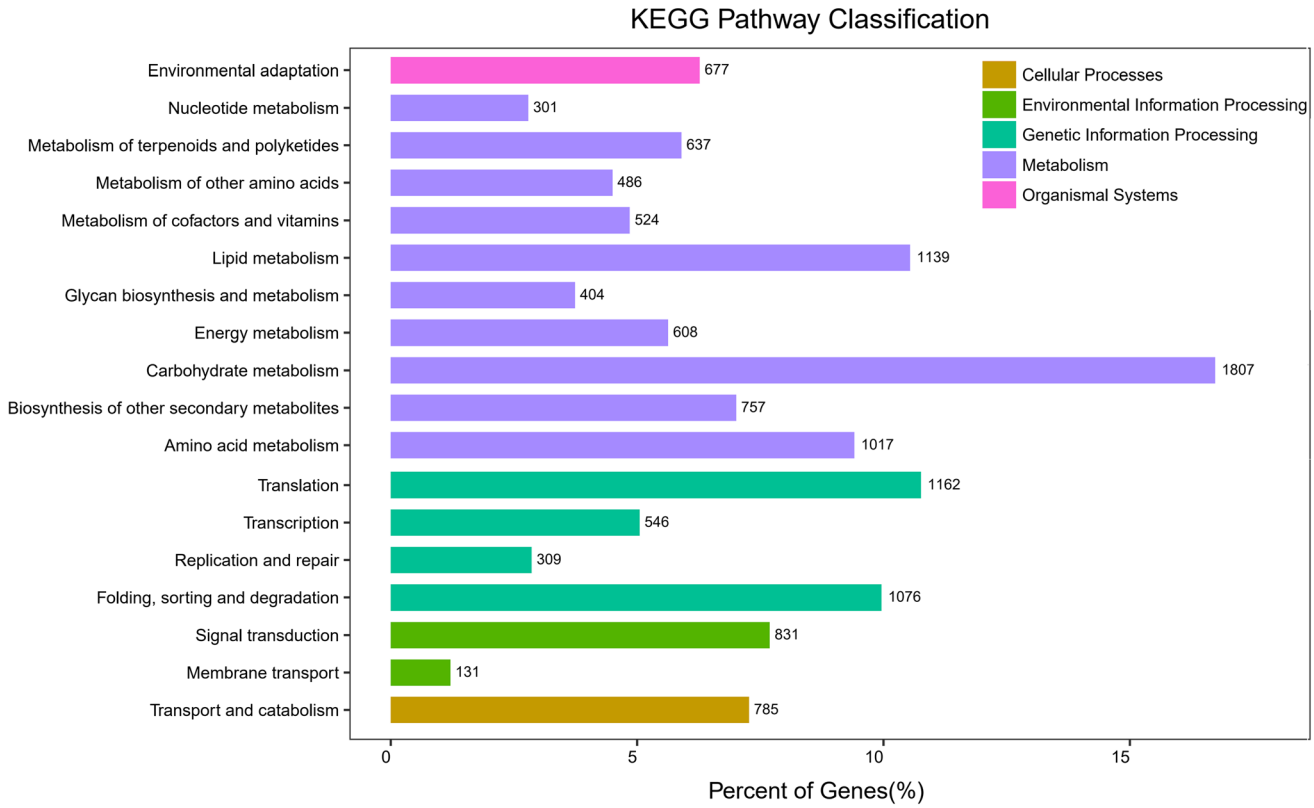

Figure 6. KEGG metabolic pathways of predicted genes in *E. cloeziana*. The *y*-axis is the name of the KEGG metabolic pathway, and the *x*-axis is the ratio of the number of genes. There are five KEGG categories including metabolism, genetic information processing, environment information processing, cellular processes, and organismal systems and their 18 sub-categories.

Table 2. Type and number of repeat motifs in SSRs of *E. cloeziana* genome.

Repeat type	Repeat motif	Number	Percentage of total SSR (%)	Percentage of each type SSR (%)	Frequency (%)
Dinucleotide	AG/CT	26850	45.64	74.51	2.70
	AT/AT	7031	11.95	19.51	0.71
	AC/GT	1961	3.33	5.44	0.20
	GC/GC	194	0.33	0.54	0.02
Trinucleotide	AAG/CTT	4404	7.49	33.88	0.44
	AAT/ATT	2592	4.41	19.94	0.26
	AGG/CCT	1532	2.60	11.79	0.15
	CCG/CGG	1504	2.56	11.57	0.15
	Other types	2965	5.04	22.81	0.30
Tetranucleotide	AAAT/ATTT	1085	1.84	44.85	0.11
	AAAG/CTTT	277	0.47	11.45	0.03
	AGGG/CCCT	266	0.45	11.00	0.03
	Other types	791	1.34	32.70	0.08
Pentanucleotide	AAAAT/ATTTT	167	0.28	21.41	0.02
	AAAAG/CTTTT	135	0.23	17.31	0.01
	Other types	478	0.81	61.28	0.05
Hexanucleotide	All types	380	0.65	100.00	0.04
Compound-SSR	All types	6220	10.57	100.00	0.63
Total		58832	100.00	–	5.91

subgenus *Idiogene*. A phylogenetic study exhibited *E. cloeziana* formed a separate clade (*Idiogene* clade) and closely related with *E. pauciflora* (*Primitiva* + *Eucalyptus* clade) (Steane *et al*. 2011). It is obvious that the genome size varies considerably among different species in eucalypts and they are consistent with their positions in phylogenies.

The genome of plants can be divided into the highly repetitive genome (≥ 50%) and lowly repetitive genomes (<

Table 3. Primer sequence and polymorphic characteristics of the 13 microsatellites in *E. cloeziana.*

Locus	Motif types	Primer sequence (F, forward and R, reverse, 5′–3′)	Blastx hit description	*T*m (°C)	Allele size (bp)	*N*a	*H*o	*H*e	HW	PIC
Ecg-007	(AAT)34	F:CCGGCTCATTGTAACACTAATACC R:CATACCACTAACGTCCTAAACTTGT	Unknown	59	169–191	5	0.707	0.568	N	0.506
Ecg-008	(AAT)33	F:TCAATTGGGTCTAGGTACATGCAT R:ACCCCATGTTAGAAAGAGCTTGAT	Unknown	60	126–160	4	0.317	0.276	N	0.254
Ecg-013	(AC)44	F:CCATTGAATAGGATCACGCAAAG R:TCTCGACGAGTAGGTTCCAAATTT	Unknown	60	140–190	9	0.750	0.838	N	0.817
Ecg-016	(GA)18	F:CGTTGGAGAGTACAGTTAGCATCT R:TTCGAGGAAGAAGAAGACGACAT	Uncharacterized protein LOC120292787 (*E. grandis*)	60	164–190	9	0.756	0.833	N	0.812
Ecg-026	(CT)38	F:CTACAACCACTTCCTCACCGTC R:TTGAGAATTCGATCTAGGACGCA	Hypothetical protein EUGRSUZ_I00994 (*E. grandis*)	60	132–177	7	0.860	0.721	***	0.685
Ecg-030	(TC)35	F:CTTGCTCTTAGACTTCCCCAGATG R:GTTCGTGAGCTAAGAAAACACAC	Unknown	59	144–170	6	0.795	0.782	N	0.747
Ecg-035	(CT)32	F:GCTCTCTCTCATCAGGGATCAAAT R:TGGCGGGATGGACATATGATATA	Hypothetical protein EUGRSUZ_B02061 (*E. grandis*)	60	185–192	3	0.773	0.542	***	0.437
Ecg-039	(CT)31	F:ATGAGGATGAGGAGGAGGATGTA R:TCTTGCTCCTTGAACACTTTGAAC	Low quality protein: respiratory burst oxidase homolog protein A-like (*E. grandis*)	60	244–253	4	0.690	0.615	N	0.540
Ecg-040	(CT)31	F:CATCACGTACCCATTCATCAGTTG R:CGAATTGCCCGACCCTATCATAT	uncharacterized protein LOC108958275 (*E. grandis*)	60	273–294	5	1.000	0.650	***	0.586
Ecg-044	(CA)29	F:CATTAAACAGATCAATTCCGACCG R:ACTGTTTACACCGTTTTAGATGCC	Disease resistance protein RPS2-like (*E. grandis*)	60	294–316	9	0.750	0.782	***	0.761
Ecg-047	(TC)29	F:GGAAATTAAGTCAATGTTGCACTG R:CTGCCCCAAGATGATAAGTTTGAC	Unknown	59	245–254	3	0.651	0.482	N	0.424
Ecg-048	(GA)16	F:AGGTCATAGCCATCATCATGTCAT R:ATGGACTAAGAATATGCGTTCCCA	Unknown	60	225–249	9	0.585	0.781	***	0.755
Ecg-050	(AT)12	F:GTGTAAATAGTGGCCCTTGTCATG R:GCCTGGTCTGTCACTATCTAGATG	Unknown	58	275–290	8	0.455	0.829	N	0.806

*N*a, number of alleles; *H*e, expected heterozygosity; *H*o, observed heterozygosity; PIC, polymorphic information content; HW, a significant departure from the Hardy–Weinberg equilibrium. *** $P < 0.001$; N, not significant.

50%), high heterozygosity (≥ 0.8%) and low heterozygosity (0.5%–0.8%) (Wu *et al*. 2014). We found the *E. cloeziana* genome contains 40.74% of repetitive elements, which is similar to the repeat content of *E. grandis* (41.22%) (Myburg *et al*. 2014) and *E. pauciflora* (44.77%) (Wang *et al*. 2020). The high heterozygosity (1.23%) for the sequencing sample was probably due to a high natural outcrossing rate in eucalypts, which are preferentially outcrossing with late-acting post-zygotic self-incompatibility resulting in outcrossing rates that can exceed 90% (Byrne 2008). Regions with high heterozygosity rate (>1) are highly problematic for *de novo* assembly (Edwards and Henry 2011; Feuillet *et al*. 2011). As expected, the quality of our pre-assembly genome was poor and only 26.1% of the region had complete gene coverage. A large number of genome regions were filtered or lost. Similar result was reported in *Betula platyphylla* (25.1%) by next-generation sequencing (Wang *et al*. 2019). Once a pair of allelic sequences exceeds a certain threshold of nucleotide diversity, these regions will be assembled as separate contigs, rather than the expected single haplotype-fused contig (Pryszcz *et al*. 2014). This results in an assembly that is significantly larger than truly haploid genome size. In the past, some traditional approaches including self-crossing, inbreeding or haploid breeding were applied to solve the impact of heterozygosity, but they are impractical for sequencing forest trees with high heterozygosity. Currently, the third-generation sequencing combined with *de novo* assembly with trio binning (Koren *et al*. 2018) and a series of new haplotype assembly software developed (Huang *et al*. 2017; Roach *et al*. 2018) make it possible.

The number of genes predicted by the genome survey only of *E. cloeziana* was much higher than that of *E. camaldulensis* (Hirakawa *et al*. 2011) genomic sequences with 3920 genes mapped onto 128 pathways, but lower than that of other sequenced genomes such as *E. grandis* (Myburg *et al*. 2014) with 36,376 predicted protein-coding genes and other Myrtaceae species like *Leptospermum scoparium* (Thrimawithana *et al*. 2019) with 31,220 protein-coding genes. The reason should be the insufficient sequence depth coverage (figure 3) and a higher proportion of the missing BUSCOs (46.5%) (figure 4). However, after combining with the transcriptome data from *E. cloeziana* terminal buds (Lan *et al*. 2021) and root (Zhu *et al*. 2018), the number of GO annotation functions and KEGG metabolic pathways were increased significantly.

Microsatellite DNA or short tandem repeat (STR) sequences are important for adaptive evolution in dynamically changing environments (Hanada *et al*. 2008). SSRs in plant genomes have been surveyed in many tree species, and the numbers were quite different among them. The occurrence frequency of SSR (5.91%) was close to that detected in *E. gunnii* (4.6%) and *E. tereticornis* (8.3%) (Vekemans and Hardy 2004), and lower than the results (25.6%) based on *Eucalyptus* genomes from GenBank reported by Dodd *et al*. (2012) and Rabello *et al*. (2005). However, the SSR search criteria in the latter included mononucleotide. Even if the same method was used for SSR analysis, the obtained distribution frequency would be different, which was related to the depth of sequencing data, the quality of sequence splicing data, the different SSR searching software and searching criteria, in addition to the inter-species differences (Brzyski *et al*. 2018). In the *E. cloeziana* genome, AAG/CTT was the most abundant trinucleotide, which was consistent with that detected in the genome of *E. camaldulensis* (Hirakawa *et al*. 2011). Of the 50 genomic SSR primers randomly selected for initial screening test, 13 primer pairs exhibited informative polymorphism in *E. cloeziana* (average number of alleles per locus, 6.23; average observed heterozygosity, 0.699). This figure was somewhat smaller than those reported by Bittencourt and Sebbenn (2007) in *E. globulus* (mean *N*a = 17.8) or by Nevill *et al*. (2013) in *E. victrix* (mean *N*a = 11). Our slightly lower values of *N*a can be explained by the relatively small sample size (44 individuals) that originated from few populations, in contrast to the entire geographic range in the other two studies.

Acknowledgements

This work was supported by the Key Research and Development Program of Guangxi Zhuang Autonomous Region (2021JBGS003) and the scientific projects of Guangxi University (202100636).

Authors' contributions

Conceptualization: W-XJ, T-DB, X-YL; methodology: X-YL; formal analysis and investigation: X-YL; writing original draft preparation: X-YL; writing review and editing: W-XJ, T-DB; funding acquisition: W-XJ; resources: J-ZW; supervision: W-XJ.

References

Altschul S. F., Madden T. L., Schaffer A. A., Zhang J., Zhang Z., Miller W. *et al*. 1997 Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3444.

Bittencourt J. V. M. and Sebbenn A. M. 2007 Patterns of pollen and seed dispersal in a small, fragmented population of the wind-pollinated tree *Araucaria angustifolia* in southern Brazil. *Heredity* **99**, 580–591.

Brooker M. I. H. 2000 A new classification of the genus *Eucalyptus* L'Hér. (Myrtaceae). *Aust. Syst. Bot.* **13**, 79–148.

Brzyski J. R., Stieha C. R. and Nicholas Mcletchie D. 2018 The impact of asexual and sexual reproduction in spatial genetic structure within and between populations of the dioecious plant *Marchantia inflexa* (Marchantiaceae). *Ann. Bot.* **122**, 993–1003.

Butrinowski R., Butrinowski I., Dos Santos E., Picolotto P., Picolotto R. and Santos R. 2013 Water availability in initial development of seedlings in protected environment *Eucalyptus grandis*. *Acta Iguazu* **2**, 84–93.

Byrne M. 2008 Phylogeny, diversity and evolution of eucalypts. In *Plant genome: biodiversity and evolution* (ed. A. K. Sharma and A. Sharma), Part E: Phanerogams-Angiosperm, vol 1, pp. 303–346. Science Publishers, Enfield.

Chor B., Horn D., Goldman N., Levy Y. and Massingham T. 2009 Genomic DNA k-mer spectra: models and modalities. *Genome Biol.* **10**, R108.

Deng Z., Chen J., Guo D., Li C. and Lu C. 2019 Genetic diversity of *Eucalyptus cloeziana*. *For. Res.* **32**, 41–46.

Dickinson G. R., Leggate W., Bristow M., Nester M. and Lewty M. J. 2000 Thinning and pruning to maximise yields of high value timber products from tropical and sub-tropical hardwood plantations. In *Opportunities for the new millennium proceedings of the Australian Forest Growers Biennial conference* (eds. A. Snell and S. Vize), pp. 32–42. Australian Forest Growers Association, Cairns.

Dodd R. S., Mayer W., Nettel A. and Afzal-Rafii Z. 2012 Clonal growth and fine-scale genetic structure in tanoak (*Notholithocarpus densiflorus*: Fagaceae). *J. Hered.* **104**, 105–114.

Edwards M. A. and Henry R. J. 2011 DNA sequencing methods contributing to new directions in cereal research. *J. Cereal Sci.* **54**, 395–400.

FAO 1979 *Eucalyptus for planting*, FAO, Rome, Italy.

Feuillet C., Leach J. E., Rogers J., Schnable P. S. and Eversole K. 2011 Crop genome sequencing: lessons and rationales. *Trends Plant Sci.* **16**, 77–88.

Grattapaglia D. and Bradshaw H. D. 1994 Nuclear DNA content of commercially important *Eucalyptus* species and hybrids. *Can. J. for. Res.* **24**, 1074–1078.

Hanada K., Zou C., Lehti-Shiu M. D., Shinozaki K. and Shiu S. H. 2008 Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiol.* **148**, 993–1003.

Heras J., Domínguez C., Mata E., Pascual V., Lozano C., Torres C. *et al*. 2015 GelJ–a tool for analyzing DNA fingerprint gel images. *BMC Bioinformatics* **16**, 270.

Hirakawa H., Nakamura Y., Kaneko T., Isobe S., Sakai H., Kato T. *et al*. 2011 Survey of the genetic information carried in the genome of *Eucalyptus camaldulensis*. *Plant Biotechnol.* **28**, 471–480.

Huang S., Kang M. and Xu A. 2017 HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics* **33**, 2577–2579.

Huang Z., Zhang J., Chen Z., Wang L. and Guo H. 2018 Development and prospects of heredity and breeding researches on *Eucalyptus cloeziana*. *J. Sic. for. Sci. Technol.* **39**, 17–21.

Koren S., Rhie A., Walenz B. P., Dilthey A., Bickhart D. M., Kingan S. B. *et al*. 2018 De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **36**, 1174–1182.

Lan J., Jiang W., Zhang L., Liang X. and Bai T. 2021 Transcriptome sequencing and bioinformatic analysis of *Eucalyptus cloeziana* terminal buds. *Mol. Plant Breeding* **19**, 5342–5351.

Leitch I. J. and Leitch A. R. 2013 Genome size diversity and evolution in land plants. In *Plant genome diversity Volume 2: Physical structure, behaviour and evolution of plant genomes* (eds. J. Greilhuber, J. Dolezel and J. F. Wendel), pp. 307–322. Springer Vienna, Vienna.

Li C. R., Xiang D. Y., Chen J. B., Zhai X. C., Kan R. F. and Lan J. 2012 Study on wood basic density variation of *Eucalyptus cloziana*. *J. Cent. South Univ. for. Technol.* **32**, 158–163.

Lv J., Li C., Zhou C., Chen J., Lia F., Weng Q. *et al*. 2020 Genetic diversity analysis of a breeding population of *Eucalyptus cloeziana* F. Muell. (Myrtaceae) and extraction of a core germplasm collection using microsatellite markers. *Ind. Crop. Prod.* **145**, 112157.

Marcais G. and Kingsford C. 2011 A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770.

Marques O. G., Andrade H. B. and Ramalho M. A. P. 1996 Assessment of the early selection efficiency in *Eucalyptus cloeziana* F. Muell. in the northwest of Minas Gerais state (Brazil). *Silvae Genet.* **45**, 359–361.

Morozova O. and Marra M. A. 2008 Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**, 255–264.

Myburg A. A., Grattapaglia D., Tuskan G. A., Hellsten U., Hayes R. D., Grimwood J. *et al*. 2014 The genome of *Eucalyptus grandis*. *Nature* **510**, 356–362.

Nevill P. G., Bradbury D., Jørgensen T., Krauss S., Samaraweera S. and Gardner M. G. 2013 Microsatellite primers identified by 454 sequencing in the floodplain tree species *Eucalyptus victrix* (Myrtaceae). *Appl. Plant Sci.* **1**, 1200402.

Nowak R. M. 2015 Assembly of repetitive regions using next-generation sequencing data. *Biocybern. Biomed. Eng.* **35**, 276–283.

Pérez-Vega E., Pañeda A., Rodríguez-Suárez C., Campa A., Giraldez R. and Ferreira J. J. 2010 Mapping of QTLs for morpho-agronomic and seed quality traits in a RIL population of common bean (*Phaseolus vulgaris* L.). *Theor. Appl. Genet.* **120**, 1367–1380.

Paiva J. A., Prat E., Vautrin S., Santos M. D., San-Clemente H., Brommonschenkel S. *et al*. 2011 Advancing *Eucalyptus* genomics: identification and sequencing of lignin biosynthesis genes from deep-coverage BAC libraries. *BMC Genomics* **12**, 137.

Peakall R. and Smouse P. E. 2012 GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. *Bioinformatics* **28**, 2537–2539.

Pryszcz L. P., Nemeth T., Gacser A. and Gabaldon T. 2014 Genome comparison of *Candida orthopsilosis* clinical strains reveals the existence of hybrids between two distinct subspecies. *Genome Biol. Evol.* **6**, 1069–1078.

Qi S. X. 2002 *Eucalyptus in China*, 2nd edition., Chinese Forestry Press, Beijing, China.

Rabello E., Souza A. N. D., Saito D., Tsai S. M. J. G. and Biology M. 2005 In silico characterization of microsatellites in *Eucalyptus* spp.: abundance, length variation and transposon associations. *Genet. Mol. Biol.* **28**, 582–588.

Ray S. and Satya P. 2014 Next generation sequencing technologies for next generation plant breeding. *Front. Plant. Sci.* **5**, 367–367.

Roach M. J., Schmidt S. A. and Borneman A. R. 2018 Purge haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 1–10.

Schmid B. and Harper J. L. 1985 Clonal growth in grassland perennials: I. Density and pattern-dependent competition between plants with different growth forms. *J. Ecol.* **73**, 793–808.

Steane D. A., Nicolle D. and Sansaloni C. P. 2011 Population genetic analysis and phylogeny reconstruction in *Eucalyptus* (Myrtaceae) using high-throughput, genome-wide genotyping. *Mol. Phylogenet. Evol.* **59**, 206–224.

Thrimawithana A. H., Jones D., Hilario E., Grierson E., Ngo H. M., Liachko I. *et al*. 2019 A whole genome assembly of Leptospermum scoparium (Myrtaceae) for mānuka research. *New Zeal. J. Crop Hort.* **47**, 233–260.

Vekemans X. and Hardy O. J. 2004 New insights from fine-scale spatial genetic structure analyses in plant populations. *Mol. Ecol.* **13**, 921–935.

Wang J. Z., Xiong T., Zhang L., Li Q. W., Fei X. Y., Shi Q. *et al*. 2016 Genetic variation and selection on growth and stem form quality traits of 25-year-old *Eucalyptus cloeziana* provenance. *For. Res.* **29**, 705–713.

Wang W., Das A., Kainer D., Schalamun M., Morales-Suarez A., Schwessinger B. *et al*. 2020 The draft nuclear genome assembly of *Eucalyptus pauciflora*: a pipeline for comparing de novo assemblies. *Gigascience* **9**, 1–12.

Wang S., Chen S., Liu C., Liu Y. and Qu G. Z. 2019 Genome survey sequencing of *Betula platyphylla*. *Forests* **10**, 826.

Waterhouse R. M., Seppey M., a Simão Felipe, Manni Mosè, Ioannidis Panagiotis, Klioutchnikov Guennadi *et al*. 2018 BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* **35**, 543–548.

Wu T. D. and Watanabe C. K. 2005 GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859.

Wu Y. F., Xiao F. M., Xu H. N., Zhang T. and Jiang X. M. 2014 Genome survey in *Cinnamomum camphora*(L.) Presl. *J. Plant Genet. Res.* **15**, 150–153.

Xie Y., Wu G., Tang J., Luo R., Patterson J., Liu S. *et al*. 2014 SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**, 1660–1666.

Xu N.-N., Jiang K., Tong X., Wang R. and Chen X.-Y. 2019 Clone configuration and spatial genetic structure of two *Halophila ovalis* populations with contrasting internode lengths. *Front. Ecol. Evol.* **7**, 170.

Zhou W., Li B., Li L., Ma W., Liu Y., Feng S. *et al*. 2018 Genome survey sequencing of *Dioscorea zingiberensis*. *Genome* **61**, 567–574.

Zhu L., Guo L., He S., Hui L., Liu X. and Chen S. 2018 Transcriptome characterization analysis of *Eucalyptus cloeziana* root based on Illumina HiSeq 2000 sequencing technology. *Mol. Plant Breed.* **16**, 4245–4254.

Corresponding editor: Shrish Tiwari