Research paper

Development of Simple Sequence Repeat Markers in *Cinnamomum kanehirae* Hayata Using Illumina-Based Sequencing

Chia-Chen Wu,[1,2] Fang-Hua Chu,[2] Cheng-Kuen Ho,[1]
Jung-Ming Chang,[1] Shu-Hwa Chang[1,3]

【Summary】

Cinnamomum kanehirae Hayata is an endemic tree species known as one of the 5 most precious softwood species in Taiwan. In recent decades, the demand for a medicinal fungus, *Antrodia cinnamomea* T.T. Chang et W.N. Chou has made the wood of *C. kanehirae* increasingly expensive. Natural stands of *C. kanehirae* are decreasing due to illegal logging. For the conservation and molecular identification of *C. kanehirae*, molecular markers are required. In this study, Illumina paired-end sequencing was used to construct *C. kanehirae* transcriptome data. In total, 58,950 unigenes were gained with an N50 of 1,292 bp after *de novo* assembly. In total, 20,135 simple sequence repeats (SSRs) were detected, among which 9,353 SSRs were successfully designed for primer pairs. In the result, the trinucleotide SSR was the most frequent motif. To date, 301 SSR primer pairs have been validated in 8 *C. kahehirae* individuals, and 94 SSR primer pairs could successfully amplify target DNA. This research generated rich genomic SSR primer information, which will be valuable with molecular identification, population genetics, conservation, and breeding studies in *C. kanehirae* in the future.

Key words: *Cinnamomum kanehirae*, *de novo* transcriptome assembly, Lauraceae, next generation sequence, simple sequence repeat (microsatellite).

Wu CC, Chu FH, Ho CK, Chang JM, Chang SH. 2018. Development of simple sequence repeat markers in *Cinnamomum kanehirae* Hayata using illumina-based sequencing. Taiwan J For Sci 33(3):197-211.

[1] Silviculture Division, Taiwan Forestry Research Institute, 53 Nanhai Rd., Taipei 10066, Taiwan. 林業試驗所育林組，10066台北市南海路53號。

[2] School of Forestry and Resource Conservation, National Taiwan University, 1 Roosevelt Road, See. 4, Taipei 10066, Taiwan. 國立台灣大學森林環境暨資源學系，106台北市大安區羅斯福路4段1號。

[3] Corresponding author, e-mail:shchang@tfri.gov.tw 通訊作者。

Received March 2018, Accepted May 2018. 2018年3月送審 2018年5月通過。

研究簡報

利用Illumina定序平台開發牛樟微衛星體分子標記

吳家禎[1,2) 曲芳華[2) 何政坤[1) 張鎔敏[1) 張淑華[1,3)

摘 要

牛樟(*Cinnamomum kanehirae* Hayata)為台灣特有種，並且被歸為台灣闊葉樹五木之一，近年來，因著牛樟芝(*Antrodia cinnamomea*)在市場上的需求提升，使得牛樟木材的需求與價格也跟著提高，而天然牛樟木也面臨許多盜伐的威脅。為了牛樟保育以及分子鑑定之工作，良好且完整的牛樟基因資料庫與分子標誌系統必須被建立。本研究使用高通量的Illumina Hi-Seq 2000定序平台建立牛樟轉錄組基因資料庫，最後透過生物資訊組裝得到58,950條unigenes，N50為1292 bp，總共搜尋到具有20,135個微衛星體序列區域，其中9,353個微衛星體序列可以成功設計引子對。這些微衛星體序列資料中，以三核苷酸重複佔最多的數量，並且使用8個不同的牛樟個體對301組微衛星體體分子標誌進行測試，其中94組可以成功擴增出目標片段。這些分子標誌可以做為未來牛樟分子鑑定、族群遺傳、保育與分子育種等各種領域廣泛應用。

關鍵詞：牛樟、轉錄組組裝、樟科、次世代定序、簡單序列重複(微衛星體)。

吳家禎、曲芳華、何政坤、張鎔敏、張淑華。2018。利用Illumina定序平台開發牛樟微衛星體分子標記。台灣林業科學33(3):197-211。

INTRODUCTION

The Lauraceae is an economically important plant family. *Cinnamomum* is one of the genera in this family and is composed of about 250 species of tropical and subtropical regions in eastern Asia, Australia, and Pacific islands. *Cinnamomum kanehirae* Hayata, an endemic and highly valuable tree species in Taiwan, is renowned for its natural properties. The wood of *C. kanehirae* has high value for its aroma, color, vein texture, and strong structure and is used in sculpture and precious furniture. *C. kanehirae* wood is in demand because of its high market price for culturing *Antrodia cinnamomea*, an endemic fungus in Taiwan, which naturally and specifically parasitizes *C. kanehirae. A. cinnamomea* is well-known for its use as a traditional Chinese medicine in Taiwanese aboriginal tribes. Nowadays, *A. cinnamomea* is in increasing demand all over the world and is well known as a herbal medicine with pharmacological effects, such as anti-cancer, anti-Inflammatory, anti-oxidant, and hepatoprotective, antihypertensive effects, anti-hepatitis B virus replication, and so on (Tzeng and Geethangili, 2011). Because of increasing demands for culturing *A. cinnamomea*, natural populations of *C. kanehirae* are rapidly decreasing in the wild due to illegal harvesting and over-cutting (Liao et al 2010). For the high demand of *C. kanehirae* wood, the price of *C. kanehirae* wood can reach US$10,000 per ton on the market. Thus, natural stands of *C. kanehirae* in Taiwan have faced serious illegal logging in the past few decades.

To trace the origin of *C. kanehirae* woods in natural stands, 15 microsatellites were successfully used in test samples. However, more microsatellite primer sets are needed to improve the assignment accuracy (Hung et al. 2017). Maternally inherited microsatellites in the chloroplast genome of *C. kanehirae* were reported (Wu et al. 2016, 2017). However, until to now, no studies showed microsatellite analysis in *C. kanehirae* on a large scale using next -generation sequencing (NGS).

Molecular markers can be used in many applications, such as DNA fingerprinting, molecular breeding, genetic analysis, and so on. Simple sequence repeats (SSR) are known as one of the molecular markers (also called microsatellites), which have short repeat motifs in particular DNA sequences. The advantages of SSR markers are high reproducibility, polymorphism among very close varieties, co-dominance, and abundant distribution in genomes (Dutta et al. 2011). There are 2 kinds of SSR: genomic and genic SSR markers. Genic SSRs specifically focus on a transcribed region in the entire genome and have more potential with linkage to genes which associated with phenotypes. In past reports, researchers developed SSR markers from Expression Sequence Tags (ESTs) library of a certain species, so-called EST-SSR markers (Sahu et al. 2012). It is more convenient to mine SSR markers from ESTs when there is a lack of genomic sequence information.

With the progress of sequencing technology, NGS provides a great way to gain abundant sequencing data. RNA sequencing (RNA-Seq) using NGS technology has been widely applied in many plant species (Mammadov et al. 2012, Poland and Rife 2012, Huang et al. 2014,Dautt-Castro et al. 2015), and sequence information for mining genic-SSRs is a feasible strategy (Dutta et al. 2011, Zhang et al. 2012a, Dai et al. 2015, Ravishankar et al. 2015). Thus, compared to the traditional method of isolating SSR from partial genomic libraries containing SSR regions by colony hybridization, NGS is a more-economic method to gain abundant molecular biological information and mine SSR markers from any species.

In this research paper, on Illumina paired-end platform was used to gain the transcriptome of *C. kanehirae* and develop SSR markers (Fig. 1). To our understanding, this is the first study to report SSR information of *C. kanehirae* using NGS technology. This study provides a large amount of SSR data in *C. kanehirae* (even for the *Cinnamomum* genus) for further research on phylogenetics population genetics, and molecular breeding.

MATERIALS AND METHODS

Plant materials and RNA extraction

RNA was extracted from roots, leaves, and stems of T3, a 1-yr-old stem cut seedling that originated from Taitung County, southeastern Taiwan RNA was also extracted from fruits of Sn54, an 18-yr-old tree planted in Singshan Nursery, New Taipei City, northern Taiwan. Samples were collected, frozen in liquid nitrogen, and stored at -80℃ in a refrigerator until use. Total RNA of each sample was extracted with the Pine Tree method (Chang et al. 1993). RNA quality was verified using on Experion™ RNA StdSens Analysis Kit (Bio-Rad Laboratories, Hercules, CA, USA). RNA was quantified with a Qubit fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) and NanoDrop 2000 (Thermo Fisher Scientfic). Eventually, equimolar concentrations of extracted RNA from 4 different tissues were pooled for complementary DNA (cDNA) library preparation. Eight *C. kanehirae* individuals were used for SSR marker validation. All plant material information is shown in Table 1.

cDNA library construction and sequencing

Poly-A containing mRNA molecules were

purified using poly-T oligo-attached magnetic beads from a eukaryote. Fragmentation mix was added to fragments of messenger RNA (mRNA). These mRNA fragments were used as reverse-transcription templates to synthesize first-strand cDNA. Second-strand cDNA was synthesized using dNTP (dUTP was replaced by dTTP) buffer, RNaseH, and DNA polymerase I.The cDNA templates were purified using a Qiagen kit (Qiagen, city, ST, USA) followed by end repair, poly A tailing, and adaptor connection. Samples were then treated with the USER™ (Uracil-Specific Excision Reagent) enzyme to digest the antisense-strand DNA followed by a polymerase chain reaction (PCR). Finally, the library could be sequenced using IlluminaHiSeq™ 2000 (Illumina, San Diego, CA, USA) performed at Yourgene Inc. (New Taipei City, Taiwan).

de novo **assembly**

Raw sequencing data were deduced in the trim process as the first step. This process converts the quality score (Q) to an error probability. Next, for every base, a new value is calculated: 0.05-Error probability. This value is negative for low-quality bases, where the error probability is high. For every base, we calculated the running sum of this value. If the sum drops below zero, it is set to zero. The part of the sequence to be retained is between the first positive value of the running sum and the highest value of the running sum. Everything before and after this region is

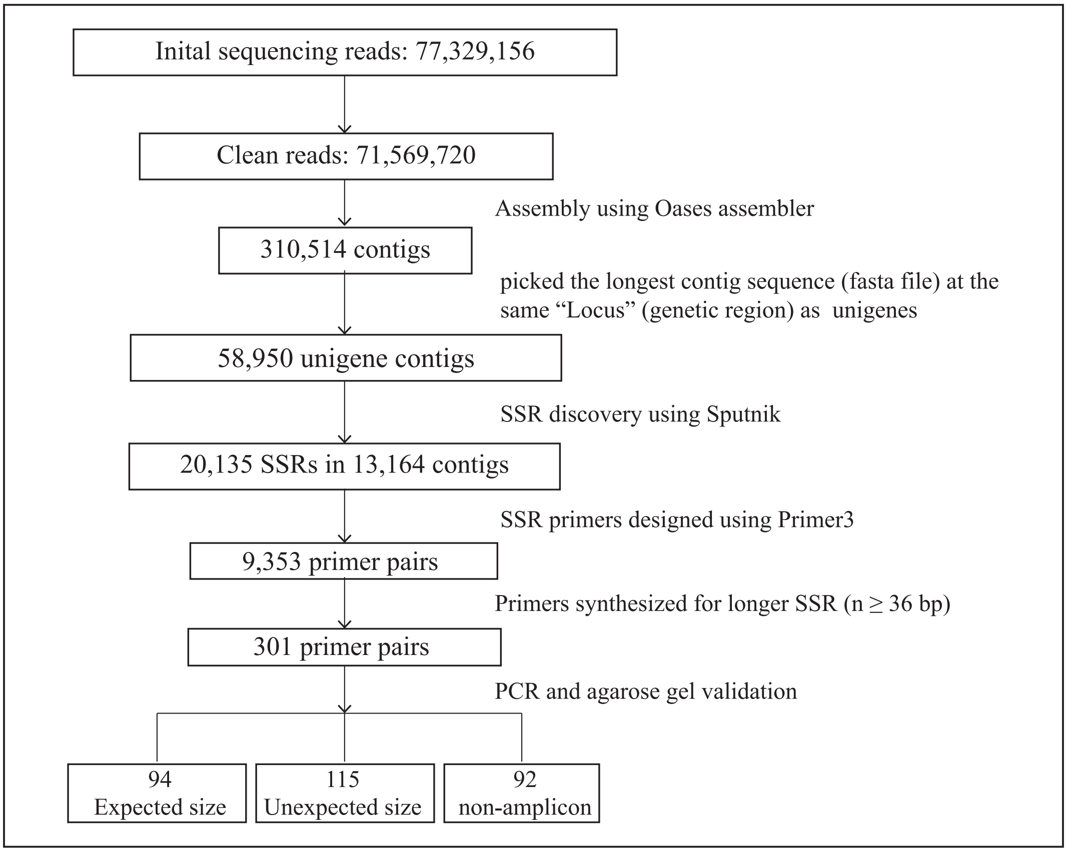
Inital sequencing reads: 77,329,156
Clean reads: 71,569,720
Assembly using Oases assembler
310,514 contigs
picked the longest contig sequence (fasta file) at the same "Locus" (genetic region) as unigenes
58,950 unigene contigs
SSR discovery using Sputnik
20,135 SSRs in 13,164 contigs
SSR primers designed using Primer3
9,353 primer pairs
Primers synthesized for longer SSR (n ≥ 36 bp)
301 primer pairs
PCR and agarose gel validation
94 Expected size
115 Unexpected size
92 non-amplicon

Fig. 1. Flow diagram of *C. kanehirae* genic-SSR marker development.

Table 1. List of plant materials

Clone name	Description	Origin	Note
T3	Roots, leaves, and stems for RNA extraction. Leaves also for DNA extraction and PCR validation	Taitung County, Taiwan	1-yr-old stem cutting seedling, cultured in a greenhouse of Taiwan Forestry Research Institute, Taipei
Sn54	Fruits for RNA extraction	Unknown	18-yr-old tree, planted in Singshan Nursery, New Taipei City, Taiwan
T4	Leaves for DNA extraction and PCR validation	Taitung County, Taiwan	Tissue-culture seedling
D1	Leaves for DNA extraction and PCR validation	Liouguei District, Kaohsiung City, Taiwan	Tissue-culture seedling
F5	Leaves for DNA extraction and PCR validation	Eastern Region of Taiwan	Tissue-culture seedling
S99	Leaves for DNA extraction and PCR validation	Forest compartment 106, Sheishankeng River Major Wildlife Habitat, Taiwan	Natural stand
LB33	Leaves for DNA extraction and PCR validation	Laonong River forest compartment 7, Pingtung, Taiwan	Natural stand
LT8	Leaves for DNA extraction and PCR validation	Chenggong forest compartment 40, Taitung, Taiwan	Natural stand
E1	Leaves for DNA extraction and PCR validation	Taimali Township, Taitung County, Taiwan	Natual stand

trimmed off. In addition, if the read length is shorter than 35 bp, the read is discarded.

To perform *de novo* assembly, all trimmed reads from samples were pooled, and Velvet (Zerbino and Birney 2008) was adopted, which uses de Bruijn graphs as the assembly algorithm and a Velvet module called Oases. It must be noted that the parameter, K-mer, is very important during *de novo* assembly and needs to be set very carefully. After the preliminary assembly produced by Velvet, we used Oases to do de novo transcriptome assembly. Oases uploads a preliminary assembly produced by Velvet, and clusters contigs into small groups, called loci. It then exploits the paired-end read and long read information, when available, to construct transcript isoforms. Transcript isoforms of a locus with the highest confidence score were selected as the unigene for that locus. In this study, these unigene were equivalent to transcripts.

SSR mining and primer design

Sputnik software (Abajian 1994) was used to find SSR motifs in unigene sequences. The program parameters of SSR motifs and minimum repeat units were set as follow: five for dinucleotide motifs, four for trinucleotide motifs, three for tetranucleotide motifs and two for pentanucleotide motifs. After finding SSRs, primer3 software was used to design PCR primers for the identified SSRs (Untergasser et al. 2012). The primer3 picks primers with a melting temperature close to 60°. The primer length is 19~22 bp. The PCR amplicon length was set to 55~550 bp.

SSR marker validation

PCR amplification of the SSR markers was performed on a Veriti 96-Well Fast Thermal Cycler (Applied Biosystems, Waltham, MA, USA) in a 30 l PCR reaction volume containing 50 ng of genomic DNA with a Taq polymerase kit (GenetBio, Daejeon, Korea). In order to reduce the cost of fluorescent primer labelling, the M13-tailed primer method was used with a minor modification (Hayden et al. 2008). Only the forward primer was synthesized with the sequence 5'- ACGACGTTGTAAAA-3'. The forward generic primer set was used with the sequence 5'-6-FAM-ACGACGTTGTAAAA-3'. We adjusted the fluorescent-labeled generic primer and reverse primer in equimolar amounts. The forward primer should be used in one-fourth the amount of the reverse primer, so that the forward generic primer can take over when the forward primer is used up. The PCR was as follows : an initial denaturation step of 5 min at 95℃, and then 30 cycles of 30 s at 95℃, 45 s at 56℃, and 45 s at 72℃, followed by 15 cycles of 30 s at 95℃, 45 s at 53℃, and 45 s at 72℃, and a final extension step of 7 min at 72℃. The PCR amplicons were validated by 2.0% agarose gel electrophoresis (Agarose-molecular biology grade, Thermo Fisher Scientfic).

RESULTS

Sequencing data and *de novo* assembly

In total, 77,329,156 raw reads with a length of 101 bp were obtained from the Illumine Hi-Seq 2000 sequencer. After trimming, the total number of paired reads was 72,180,858, with an average length of 96.9 bp. After *de novo* assembly of trimming reads, 310,514 transcripts with a total of 250,223,161 bp were obtained. The total number of unigenes yielded was 58,950 comprising 52,356,624 bp. The fragment sizes of the unigenes were 200~12,438 bp, with an average length of 888 bp and an N50 of 1292 bp (Table 2). The distribution of the unigene length is shown in Fig. 2.

SSR identification and primer design

In total, 13,164 of 59,890 unigenes contained 20,135 SSRs in our result. Of these, 9,353 (46.45%) of 20,135 SSRs were successful for designing primer pairs. In an analysis of SSR repeat types, the amount of trinucleotide repeats was most abundant (10,301) at 51.1% of total SSRs, following by dinucleotide repeats (30.1%), pentanucleotide repeats (9.5%), and tetra-nucleotide repeat (9.1%) (Fig. 3). Among the designed SSR

Table 2. Summary of *C. kanehirae* illumina-based transcriptome sequencing reads and their assembly data

Parameter	Transcriptomic data of *C. kanehirae*
Total no. of reads	77,329,156
Read length	101
Total no. of bases	7,810,244,756
Number of reads after trimming (paired)	71,569,720
Average length after trimming	96.9
Total no. of clean bases	6,935,105,868
Total no. of contigs	310,514
No. of unigenes	58,590
Average length of unigenes (bp)	888
N50 (bp)	1,292

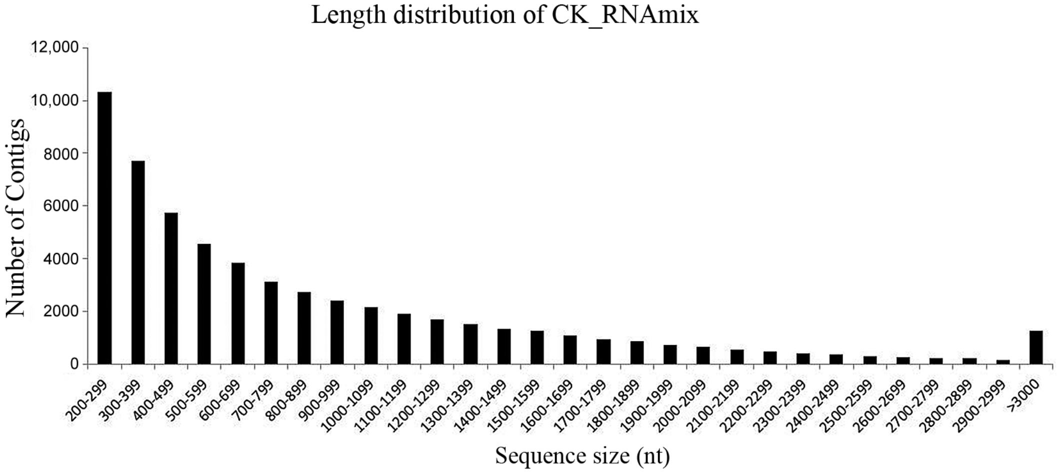

Fig. 2. Length distribution of *C. kanehirae* transcriptomic unigenes.

primers, the same results were found for the total SSR repeat type analysis, with trinucleotide repeats being most abundant (61.3%), followed by dinucleotide (23.5%), tetranucleotide (7.9%), and pentanucleotide repeats (7.2%). In the total SSR identification, the unit of the CT repeat was most common (7%), followed by TC (7%), GA (6%), and AG (5%) repeats. Among trinucleotide units, AGA/TTC/TCT/GAA were predominant (Fig. 4). In an assay of SSR repeat times, the four-time repeats were the most abundant. The highest repeat time was for a dinucleotide (CT) was 44 repeat times. The high percentage (73.49%) of repeat times was distributed in 3 to 7 time repeats (Fig. 5).

SSR marker development and validation

The top 301 primer pairs with an SSR length larger than 30 bp were selected by sorting SSR repeat lengths from 9,353 designed primer pairs. These 301 primer pairs were validated by a PCR with eight different *C. kanehirae* individuals and analyzed by gel electrophoresis. Eventually, 94 primer pairs successfully yielded PCR amplicons of an expected size (PCR success rate was 31.23% of 301 primer pairs) (Table 3). Another 115 primer pairs amplified an unexpected size product (including multiple amplicons, of a larger or smaller size) and 92 primer pairs failed to amplify any product (no PCR amplicon).

DISCUSSION

C. kanehirae is an endemic tree species in Taiwan and nowadays it has been exported and widely planted in mai China and southeast Asian countries. only fruiting bodies of *A. cinnamomea* that live on *C. kanehirae* can produce high-quality medicinal components, such as ergostane-type triterpenoids and benzolic (Lin et al 2011, Wu et al 2017), *C. kanehirae* wood and seedlings have become increasingly popular in the market. The illegal logging of *C. kanehirae* occurs frequently in Taiwan. For conservation and tracking the origin of *C. kanehirae* wood, it is necessary to develop a useful molecular marker system. Insertion/deletion markers from the complete chloroplast of *C. kanehirae* were validated to distinguish *C. kanehirae* from *C. micranthum* (Wu et al. 2017). However, more microsatellites are required to improve the identification accuracy of origins (Hung et al. 2017). In this report, we provided many potential SSRs and

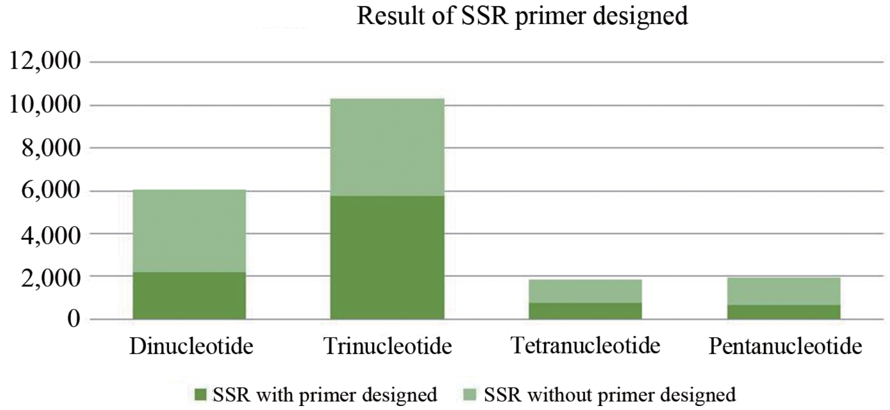

Fig. 3. Distribution of different SSR types. It indicates the amount of SSR with a primer designed (green) and of SSRs without a primer designed (light green).

validated SSR primers for *C. kanehirae* for future research.

In addition, there were no ESTs data of *C. kanehirae* and other *Cinnamomum* species deposited in NCBI before. In the present study, 310,514 transcripts and 58,950 unigenes were acquired and could provide useful bioinformatics resources on gene construction and other molecular research in *C. kanehirae*, and even for the *Cinnamomum* genus.

SSRs are widely used in crop breeding programs and in population genetics because they are reliable, polymorphic, reproductive, codominant, and easy to amplify. However, in non-model species like *C. kanehirae*, there is a lack of available genomic resources. Until now, only 15 nuclear SSR markers for *C. kanehirae* had been published (Hung et al. 2014). In the past, the method of SSR marker development was expensive and time-consuming, involving construction and screening of SSR-enriched genomic DNA libraries. The advent of NGS has made SSR discovery easier and less complicated. NGS technologies have become useful and powerful tools with relatively inexpensive costs to rapidly gain massive sequence data on many organisms (Wu et al. 2014, Dai et al. 2015, Han et al. 2015;). The work flow in this report is suitable for many tree species for which no or only a few SSR markers are available.

The assembly unigenes with a total of 52,356,624 bp cover 52.3 M bp of the *C. kanihriae* genome. This is approximately 7.6% of the *C. kanehirae* genome, which has been measured by flow cytometry to contain 688 M bp of DNA (Wu et al. 2015). In total, 20,135 SSRs were identified in 58,950 unigenes, with a mean of one SSR per 2.6 kb. An older report mentioned that the average frequency of SSRs was one per 6.04 kb in genomic DNA, decreasing to one per 14 kb in ESTs (Cardle et al. 2000). In recent reports, the density of *C. kanehirae* SSRs was relatively higher than in other crops, such as wheat (one per 28.32 kb), rice (one per 11.81 kb), maize (one per 28.32 kb) and soybeans (one per 23.80 kb) (Zhang et al. 2012a); but similar to other woody plants, including tea (one per 3.98 kb) (Ma et al. 2014), coffee (one per 2.16 kb) and *Populus tomentosa* (one per 8.24 kb) (Du et al. 2012). However, detection of the SSR frequency is correlated with many factors, such as software detection criteria, data size, mining software, and different species and materials.

The total of 20,135 SSRs was obtained from 13,164 unigenes (22.33%), and 4,542

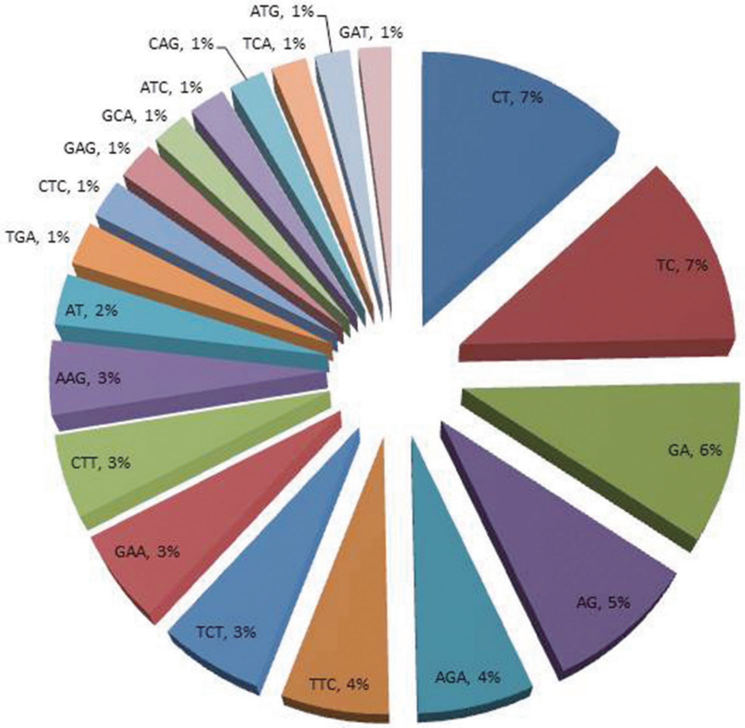

Fig. 4. Ratios of the top 20 SSR units in *C. kanehirae* genic-SSR.

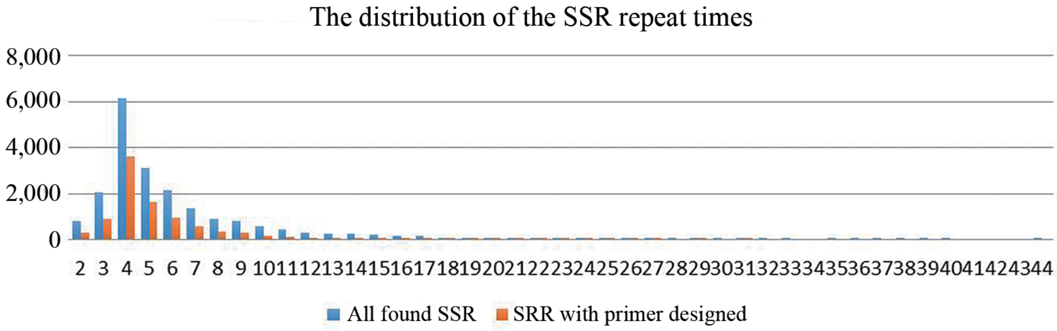

Fig. 5. Repeat time distribution of *C. kanehirae* genic-SSR.

of the 13,164 unigenes contained more than one SSR. The trinucleotide repeat motif was the most abundant, which is concordant with Ma bamboo (*Dendrocalamus latiflorus*), mango (*Mangifera indica*), oil palm (*Elaeis guineensis*), *Pelargonium* spp. and *Camelina sativa* (Liu et al. 2012, Mudalkar et al. 2014, Xiao et al. 2014, Ravishankar et al. 2015, Narnoliya et al. 2017). In many species, most of the SSR markers detected are dinucleotide repeats, but these are less frequent in coding regions. Trinucleotide repeats are more common in coding regions, because they do not cause a frameshift (Merritt et al. 2015). Thus, it is reasonable that the most abundant SSRs in *C. kanehirae* RNA-Seq were trinucleotide repeats.

In our results, the predominant SSR

Table 3. Successfully amplified genic-SSR markers in *C. kanehirae* transcriptome

No.	Primer name	Size of PCR product (bp)	SSR type	SSR length (bp)	SSR motif	Forward primer	Reverse primer
1	CK-SSR1	208	dinucleotide	60	GA	GTCCTGTGCATCGATCTAGGG	CAGCTCCCTCTCCTTCCCT
2	CK-SSR2	407	dinucleotide	60	GA	AGGACGGACAGCTAAAGTGC	ATGTGGTCCACATGAGACCAC
3	CK-SSR3	241	trinucleotide	60	CCA	CCATCCACAGATGTCCCAACA	ATTTGCAGGAGTTGGAGGAGG
4	CK-SSR7	231	tetranucleotide	32	CTTC	CATCCTTCACTCTCTCTCGCTC	GAGCAGAACCGAGAGTGAAAGA
5	CK-SSR8	251	pentanucleotide	35	CTTCC	ACCAACTCCAATTCAACACCAA	AAAGCGCGCAAAAATGTCTCT
6	CK-SSR10	219	dinucleotide	36	GA	CCATTCCGAGAGACCGTTCC	AGGATGTGCTTGAGGAGTGTG
7	CK-SSR11	265	dinucleotide	52	TC	TAAGCTTTGTCCTTGCGTTGC	TCGAAACCTGCGAGAACATGA
8	CK-SSR13	318	trinucleotide	51	GAA	TCCATACCCAACGACAAGGTG	CGACTGGAGTAACCTGCACTT
9	CK-SSR20	366	dinucleotide	48	TC	TGCACCGTCATTTCCAGATGA	GGGATTTGCAACTTGGTCCAAA
10	CK-SSR21	361	dinucleotide	46	CT	CACAACCACACCTCCTCTCTC	CACACCCAAATCCCAACACATC
11	CK-SSR26	310	dinucleotide	44	CT	CTTGCCCTTGGACTGGGAG	TTCTAAACACCACCCCGCATAA
12	CK-SSR29	375	trinucleotide	42	AAG	GATTGGTGATGGTCCTGTTCCT	ACTGACCTTGAGAAGTCTGCAC
13	CK-SSR32	371	trinucleotide	42	AGA	CAACTGTCCATAGGGTTCAGCT	GCGCATGGATTTTCGTCTCTC
14	CK-SSR38	409	dinucleotide	42	TC	CTCACTCCACCTTCCAACCTC	AAAGGTGGGCTGATCTGAAGAG
15	CK-SSR45	223	dinucleotide	40	CT	GCTTCAGAGTTGCAACACCAAT	TGGCCCGAGTTAGAAATTGCT
16	CK-SSR49	213	dinucleotide	40	GA	TTGCAGATCATCCCAACTGCT	CAAAACCAAACAACCACCCCAT
17	CK-SSR52	400	dinucleotide	40	GA	GCAATTTCATCGTCTCCGGC	TCCTCCTCTTCACCCCTCG
18	CK-SSR57	173	trinucleotide	39	GCA	GCAATCGCAACAGCCACAA	CACTGCCAACAAGACCATTGG
19	CK-SSR60	144	dinucleotide	38	CT	CCAGGTTCCACTCGGATGAAA	CCATCAATGGCAAACCCCATTT
20	CK-SSR64	277	dinucleotide	38	GA	ATACACAGATAGCCATGCCGG	AATTCCGCAAGAACGACAGTTG
21	CK-SSR70	236	dinucleotide	38	TC	TCCATAGCCTTCCTGTCCTCT	AAAATGGCGGCAACTCAACTC
22	CK-SSR72	418	dinucleotide	38	TC	TGTTTGCGTTGCAAGCTATTCA	GGTGCGATTCCCCTTTGAGA
23	CK-SSR73	281	dinucleotide	38	TC	TGCAAACTAGTGTTTCTCCCAT	TTGGACAACGCTGAGATGATGA
24	CK-SSR74	193	trinucleotide	36	AAG	TCCAAGGTTCTTCGAGCTTCC	TCTGTCGAGTTTGTACGGTGG
25	CK-SSR80	323	trinucleotide	36	CAG	CCCAGGTTTTGTAATGCCCTTG	GATCAAACCGCTGTTGCTGTT
26	CK-SSR81	455	trinucleotide	36	CCT	GAAGACTGCAGCAGCTCCA	CCATGACTTCGCACACATGTTT
27	CK-SSR82	158	dinucleotide	36	CT	AGGTTTGTTCTCTAGTGGGCTG	TAGTCAATCACACACGCACGA
28	CK-SSR85	100	dinucleotide	36	CT	CAACATGTCCATGTACCTGAGA	AAATGGGAAGAACGGCCACTAT
29	CK-SSR86	306	dinucleotide	36	CT	AACACCATTCATGCTGCCAAC	TCGATCACCATCATCATCGTCC
30	CK-SSR87	341	dinucleotide	36	CT	AGTGTCGCAGTCTTCAGGC	GGCCTGGCCAAATACTCAAATC
31	CK-SSR88	99	dinucleotide	36	CT	TCTTTCACATCTTTCCCTCACA	AGGCCATGTCTCTCTCCGT
32	CK-SSR89	323	trinucleotide	36	CTG	GATCAAACCGCTGTTGCTGTT	CCCAGGTTTTGTAATGCCCTTG
33	CK-SSR91	385	dinucleotide	36	TA	AGTTCTAAGCTCTGCCCCAAC	TTCTTAGCCGTCCATTGCTGT
34	CK-SSR93	100	dinucleotide	36	TC	GTGCTCAAGCTCTATGATTCCA	GAAAGCAGGCCACAAGCAC
35	CK-SSR98	228	dinucleotide	36	TC	TCCTTCAACCAGTTCCTTACAA	TGGTTTCTGATAGGCCTCAACC
36	CK-SSR99	294	trinucleotide	36	TCA	ACCCTTCTTGTCGTTCATCCC	ATGAAGCTACCCCAGTTCAAGG
37	CK-SSR103	227	dinucleotide	34	AG	GCATCGAGAGAGAGAGGGAAAG	TGACGATCATGGGCAGTGAAA
38	CK-SSR106	370	dinucleotide	34	AG	GGTGGAGGTGGTGAGCTTC	AACGAGGTTCATGCTTGCAAC
39	CK-SSR112	193	dinucleotide	34	AT	GGGTTGGGGTGTTTGTTTTGAA	CCAATTCGTTTCCAAAAGCCCT
40	CK-SSR117	130	dinucleotide	34	CT	CCCCTCATCTGATCGTTTCGT	TCTCCACCAGCCATCGGTA
41	CK-SSR120	357	dinucleotide	34	GA	TGGAGGGGTGTGGAGTATAGTT	TGTGAACGTTGAAGGGAACTCA
42	CK-SSR123	311	dinucleotide	34	GA	AGAGAGGAGCCCCAGTGAAT	CCAATTTCCAACACCACTCACC
43	CK-SSR124	395	dinucleotide	34	GA	TCAGCTCTGAAATGCCTCCATT	GCCGACAGCCTTCTTCACT
44	CK-SSR130	127	dinucleotide	34	TC	GACGACAATGGGGAAGAGAAGT	TTTTGAAAGAAAGCGCGGAGG
45	CK-SSR131	273	trinucleotide	33	AAG	CACAAGGTTTTGGCACTCTCAG	TCGCCATGCCCTTTTACTTCT

con’t

46	CK-SSR132	331	trinucleotide	33	AGA	AGAACGAGCTGCAATCGAAGA	AGGCAGAGAAGCAGAAAGGTC
47	CK-SSR135	268	trinucleotide	33	AGA	TGGCGGAGGTCATTTTCGTTA	AGAGCTCGTCCCAAAGTGC
48	CK-SSR136	229	trinucleotide	33	AGA	TTGTCTTAGATTTTGCAGCAGC	AGCTCCCATTCAAGGAAGACAG
49	CK-SSR141	161	trinucleotide	33	GAA	AATCCCTCAACAAACCCCAACT	TCCAGAACAGCAGCATCCATT
50	CK-SSR142	406	trinucleotide	33	TCT	CATGGTGTGCATCCCACCT	TCTCCCCTCACTCTCCAGC
51	CK-SSR143	355	trinucleotide	33	TCT	AAAGCTGTCATGGAGAGGACTG	AAGCAAATGGAAGCAACCGAC
52	CK-SSR145	375	trinucleotide	33	TTC	CTTCTTCCCCAACCTCTCTCAC	GAAATCAAAGCGACAGTGCAGT
53	CK-SSR146	216	trinucleotide	33	TTC	TCTTTTGCTCCCCTTCATCACA	GGGACTGACTTGGTTCGGATT
54	CK-SSR147	202	trinucleotide	33	TTC	TCTGCAAATTCCAGCAAGTGC	TCCGAACCATACGATCTTAGCG
55	CK-SSR148	255	trinucleotide	33	TTC	CCTCTGCTGCAACTCCCC	GAGAAGATTCCGAGAGCGAGG
56	CK-SSR149	308	trinucleotide	33	TTC	TACCAATACCCGCGTTGTTGA	AAGGGAGAGGCCTCGACAT
57	CK-SSR151	268	dinucleotide	32	AG	TGGCGGAGGTCATTTTCGTTA	AGAGCTCGTCCCAAAGTGC
58	CK-SSR154	329	dinucleotide	32	AG	CTATTCCTACCCCAAAGGCCTC	CCGGCCAAAAATACAGTTGCA
59	CK-SSR158	214	dinucleotide	32	AT	TTAGTGGTGCTCAATTCCGCT	CCTCACCACTCACGAGTCAC
60	CK-SSR159	362	dinucleotide	32	CT	ACCTTCATGCCCCAACGAAT	ACCGAGTTATGACACTTGGGAC
61	CK-SSR160	371	dinucleotide	32	CT	AAGAGGCAACATGGGAGCC	AAAGACCATTGCTCAGGGAGG
62	CK-SSR165	112	dinucleotide	32	CT	TCGGCTTCCATTTCTTACATCA	TCTGCAAAGTCTTCTGCGCTA
63	CK-SSR166	331	dinucleotide	32	CT	ACTCCCAGCCAGAGGAGAG	TGAGTTTGTACGGAAGGGTCG
64	CK-SSR167	409	dinucleotide	32	CT	CAGCTAAATGTGGCAAGCAAGT	CAAGGAAGCTAAGTTGCTCAGC
65	CK-SSR168	377	dinucleotide	32	CT	CTCAGGATCAAGGTGCAGAGTT	AAGAGAAGACCAGACACAGCG
66	CK-SSR169	376	dinucleotide	32	CT	GCTTGCTTCCAAAATGGGGATT	CGGTTTGGTTGGTGCAGC
67	CK-SSR173	114	dinucleotide	32	GA	GGGAGGGAGAGATTATCGAGGA	CGTCCCTCTCTCCCTTCCA
68	CK-SSR181	382	dinucleotide	32	GA	CCAGGTGCACCCTACACAG	TGGAAATCGCTTACCGTGACA
69	CK-SSR188	228	tetranucleotide	32	TATC	CCCCATCTTTCTCGAGTCTCG	AAAATTTTGGCCAGAAGACCGG
70	CK-SSR190	235	dinucleotide	32	TC	ACCATCCTGACTCCATGTTTCC	AGAGGGAACTTGGACTGCATG
71	CK-SSR193	235	dinucleotide	32	TC	CATTCCTTCCTTGTTTGCAGCA	AGAGTGGTTTCTCATTGGCCAT
72	CK-SSR195	128	dinucleotide	32	TC	ACTTGACTCCTTGACTTCAAGT	GAGCAGAGGAGAAGAGCGATC
73	CK-SSR200	322	trinucleotide	30	ACA	TGCTCCTTCACCCACTCCT	GAAGACTGTGGAGAAATCCCGT
74	CK-SSR202	343	trinucleotide	30	ACC	CCAGTTCAAGCCCCTCAGG	ATGAAAATCCATGGCCCCTGT
75	CK-SSR204	246	trinucleotide	30	ACC	ACCCTTCCAAAATCCCGTACC	TTGCCCTTCGCACTAATTCCT
76	CK-SSR221	369	trinucleotide	30	AGA	GCCGACTGCGTATTCAAATCC	CCTATGGTCCGCCTCAATCC
77	CK-SSR222	145	trinucleotide	30	AGA	ACCCTAGCCGTCGTCTTCT	CCGCGATGATTTCTTCACTGTC
78	CK-SSR223	315	trinucleotide	30	AGA	ACCCCAATTCCAATTCCCCTC	GATAAGCCATCGGAGGTGGAG
79	CK-SSR224	400	trinucleotide	30	AGA	ATGCCCATCTCGTCCATGATC	ACCTTCGACCCATTTCTCCAC
80	CK-SSR225	177	trinucleotide	30	AGA	GCGAGGCTCAACAAAACCATT	GAGCGGAACACCTTTTTGCAT
81	CK-SSR229	219	trinucleotide	30	CAA	ATGTATCAAACACTGGGCGGT	CAACCAACCAGAGGCAGAGAT
82	CK-SSR230	317	trinucleotide	30	CAA	CGGCTCCGCTCTCCAAAA	GATCGATGTTGGCGTCGTTG
83	CK-SSR231	254	dinucleotide	30	CT	AGGCATGTTTCTTCCCTGAGG	AACCTTGTGTTCTGTGCCCA
84	CK-SSR241	220	dinucleotide	30	CT	TTTCTTGCCCTAGACGTTCCTC	TGTTACCTGTTGTCACAACCCA
85	CK-SSR263	224	trinucleotide	30	GAA	GCCCCCAAATCGAAACCCTA	GCGTCCTCGTTCATTGAATCG
86	CK-SSR266	196	trinucleotide	30	GAA	TTGGCCGTCTAAGACACACTC	CCTGCTTCGATCGAGTGATGA
87	CK-SSR268	171	trinucleotide	30	GAA	ATCCCAACACGAAAGAAACCCT	CCCATGCTTCGAACGATCTTTC
88	CK-SSR271	261	dinucleotide	30	GT	GTTAAGCACGAGGGGCTGT	GCACCACTGCACAGTTTAAGG
89	CK-SSR274	295	dinucleotide	30	TC	AGAGGAAGTGACAGAATGCAGG	CTGGAAAGTCGACTGGGATTCA
90	CK-SSR284	146	dinucleotide	30	TC	CGACCGGAAGAAGCTTCTCTT	GAGGTGATTTGCAGAGGGGAA
91	CK-SSR285	126	dinucleotide	30	TC	TCCTACTCTCCTCTGTCTTGTC	TGCAGAAGACAGAGCTGGAAG
92	CK-SSR286	390	dinucleotide	30	TC	ATGGAGGGTTTTGGCGAGG	CTCATGTGTCTCGGACTCACAA
93	CK-SSR289	336	trinucleotide	30	TCT	ACGCTCATTCCAAAGCCGA	AACCACCCTCTCTCTCTCTCC
94	CK-SSR298	211	trinucleotide	30	TTC	AAGCCGTGCCCCTTTCTC	AGTTCCGATTTGGCTGCGA

units were CT/TC/GA/AG, which is concordant with most vascular plants which have frequent motifs of AG/CT and GA/TC (Victoria et al. 2011). The CT/TC units occupied about 14% of SSR in total predicted SSRs in *C. kanehirae* RNA-Seq. This may be because CT repeat units can have the function as an enhancer, which found the "TCTCTCTCT" motif downstream of the transcription start site of *CaMV* 35S RNA, which can enhance gene translation in plant protoplasts (Ge et al. 2014). In addition, the complement of $(CT)_n$, $(GA)_n$ is a regulatory element involved in light regulation (Zhang et al. 2006).

Distributions of SSR repeat times may differ according to different detected parameters. However, the SSR repeat times, length and type are correlated with the degree of polymorphism (Dutta et al. 2011, Sahu et al. 2012, Merritt et al. 2015). In our results, only 31.23% of 301 SSR primer pairs could be successfully amplified as of PCR products expected size. The success rate was lower than in other reports, sesame (92%), mango (81.81%), pigeon pea (80%), barley (67%), and sugarcane (48.5%) (Dutta et al. 2011, Zhang et al. 2012b, Ravishankar et al. 2015). This may be attributed to the sequencing coverage, depth, and primer chosen. The genome size of *C. kanehirae* is 688 M bp, and the sequencing depth in the present data was only 10.08-fold. In addition, the 301 primer pairs were selected from the longest SSR length to 30 bp. Long SSR repeats may produce faulty assembly contigs during the *de novo* assembly process. Another possible explanation for the non- or un-excepted PCR amplicons is the flanking primer pairs located at a splicing site with an intron or chimeric cDNA contig (Dutta et al. 2011). Among these 94 SSRs, dinucleotides were predominant (54), followed by trinucleotides (37), tetranucleotides (2) and a pentanucleotide (1). Choosing further SSR markers should carefully be considered for different purposes. In general, dinucleotide SSR will be potentially genetically variable, most likely due to easy DNA slippage during replication. But dinucleotide SSRs can also lead to difficulty in scoring alleles (i.e., stutter peaks, and PCR errors with slippage) compared to larger-length motifs such as penta- and hexanucleotide motifs. In addition, for studies of mating system estimation and population genetics, larger repeat times and longer-length SSR are recommended. Lower repeat times or interrupted repeats are suitable for working on a coarser scale with more distantly diverged species or taxa (Merritt et al. 2015). SSRs are useful information for discovering DNA polymorphisms between ecotypes or varieties. This research is the first report to develop and validate a comprehensive set of genic SSR markers in *C. kanehirae*, even for the *Cinnamomum* genus. 94 of 9,353 genic-SSR markers were validated and will be used for further diversity analysis as well as origin tracking, species identification, and molecular breeding in *C. kanehirae*.

CONCLUSIONS

In the present report, the first *C. kanehirae* RNA-Seq and genic-SSR data were provided. In total, 58,950 unigenes were gained with an N50 of 1,292 bp and 9,353 of 20,135 SSRs were successfully designed for primer pairs. The 301 SSR primer pairs were validated, and eventually, 94 SSR primer pairs could successfully amplify target DNA, which will be useful for further molecular biological work. The information generated in this study will provide useful molecular tools for genetic studies and breeding of *C. kanehirae*.

ACKNOWLEDGEMENTS

We thank the Ministry of Science and Technology and Taiwan Forestry Research Institute, Taiwan which supported this work and provided grants (MOST 102-2313-B-054-003-MY3 and 106AS-11.1.3-FI-G1)

LITERATURE CITED

Abajian C. 1994. Available at Sputnik. http://espressosoftware.com/pages/sputnik.jsp. Accessed dd mo year.

Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh R. 2000. Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics 156:847-54.

Chang S, Puryear J, Cairney J. 1993. A simple and efficient method for isolating RNA from pine trees. Plant Mol Biol Rep. 11:113-6. doi:10.1007/BF02670468.

Dai F, Tang C, Wang Z, Luo G, He L, Yao L. 2015. De novo assembly, gene annotation, and marker development of mulberry (*Morus atropurpurea*) transcriptome. Tree Genet Genomes 11. doi:10.1007/s11295-015-0851-4.

Dautt-Castro M, Ochoa-Leyva A, Contreras-Vergara CA, Pacheco-Sanchez MA, Casas-Flores S, Sanchez-Flores A, et al. 2015. Mango (*Mangifera indica* L.) cv. Kent fruit mesocarp *de novo* transcriptome assembly identifies gene families important for ripening. Front Plant Sci 6:62. doi:10.3389/fpls.2015.00062

Du QZ, Zhang DQ, Li BL. 2012. Development of 15 novel microsatellite markers from cellulose synthase genes in *Populus tomentosa* (Salicaceae). Am J Bot 99: 2011-3. doi:10.3732/ajb.1100308.

Dutta S, Kumawat G, Singh BP, Gupta DK, Singh S, Dogra V,et al. 2011. Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea [*Cajanus cajan* (L.) Millspaugh]. BMC Plant Biol 11:17. doi:10.1186/1471-2229-11-17.

Ge X, Chen H, Wang H, Shi A, Liu K. 2014. *De novo* assembly and annotation of *Salvia splendens* transcriptome using the illumina platform. PLoS One 9:1-9. doi:10.1371/journal.pone.0087693.

Han S, Wu Z, Jin Y, Yang W, Shi H. 2015. RNA-Seq analysis for transcriptome assembly, gene identification, and SSR mining in ginkgo (*Ginkgo biloba* L.). Tree Genet Genomes 11. doi:10.1007/s11295-015-0868-8.

Hayden MJ, Nguyen TM, Waterman A, Chalmers KJ. 2008. Multiplex-Ready PCR :A new method for multiplexed SSR and SNP genotyping. 12:1-12. doi:10.1186/1471-2164-9-80.

Huang YY, Lee CP, Fu JL, Chang BC, Matzke JM, Matzke M. 2014. *De Novo* transcriptome sequence assembly from coconut leaves and seeds with a focus on factors involved in RNA-directed DNA methylation. G3 Genes|Genomes|Genetics 4:2147-57. doi:10.1534/g3.114.013409.

Hung KH, Lin CH, Chuan SH, Chung CY, Ju LP. 2014. Development, characterization and cross-species amplification of new microsatellite primers from an endemic species *Cinnamomum kanehirae* (Lauraceae) in Taiwan. Conserv Genet Resour online ver. doi:10.1007/s12686-014-0239-z.

Hung KH, Lin CH, Ju LP, 2017. Tracking the geographical origin of timber by DNA fingerprinting:a study of the endangered species *Cinnamomum kanehirae* in Taiwan. Holzforschung. doi:10.1515/hf-2017-0026.

Liao PC, Kuo DC, Lin CC, Ho KC, Lin TP, Hwang SY. 2010. Historical spatial range expansion and a very recent bottleneck of *Cinnamomum kanehirae* Hay. (Lauraceae) in Taiwan inferred from nuclear genes. BMC Evol Biol 10:124. doi:10.1186/1471-2148-10-124.

Lin TY, Chen CY, Chien SC, Hsiao WW,

Chu FH, Li WH,et al. 2011. Metabolite profiles for *Antrodia cinnamomea* fruiting bodies harvested at different culture ages and from different wood substrates. J Agric Food Chem 59:7626-35. doi:10.1021/jf201632w.

Liu M, Qiao G, Jiang J, Yang H, Xie L, Xie J, Zhuo R. 2012. Transcriptome sequencing and de novo analysis for Ma bamboo (*Dendrocalamus latiflorus* Munro) using the Illumina platform. PLoS One 7:1-11. doi:10.1371/journal.pone.0046766.

Ma JQ, Yao MZ, Ma CL, Wang XC, Jin JQ, Wang XM, Chen L. 2014. Construction of a SSR-based genetic map and identification of QTLs for catechins content in tea plant (*Camellia sinensis*). PLoS One 9. doi:10.1371/journal.pone.0093131.

Mammadov J, Aggarwal R, Buyyarapu R, Kumpatla S. 2012. SNP markers and their impact on plant breeding. Int J Plant Genom 2012. doi:10.1155/2012/728398.

Merritt BJ, Culley TM, Avanesyan A, Stokes R, Brzyski J. 2015. An empirical review: characteristics of plant microsatellite markers that confer higher levels of genetic variation. Appl Plant Sci 3:1500025. doi:10.3732/apps.1500025.

Mudalkar S, Golla R, Ghatty S, Reddy AR. 2014. *De novo* transcriptome analysis of an imminent biofuel crop, *Camelina sativa* L. using Illumina GAIIX sequencing platform and identification of SSR markers. Plant Mol Biol 84: 159-71. doi:10.1007/s11103-013-0125-1

Narnoliya LK, Kaushal G, Singh SP, Sangwan RS. 2017. *De novo* transcriptome analysis of rose-scented geranium provides insights into the metabolic specificity of terpene and tartaric acid biosynthesis. BMC Genom 18:74. doi:10.1186/s12864-016-3437-0.

Poland JA, Rife TW. 2012. Genotyping-by-sequencing for plant breeding and genetics. Plant Genome 5:92-102. doi:DOI 10.3835/plantgenome2012.05.0005.

Ravishankar K V, Dinesh MR, Nischita P, Sandya BS. 2015. Development and characterization of microsatellite markers in mango (*Mangifera indica*) using next-generation sequencing technology and their transferability across species. Mol Breed 35. doi:10.1007/s11032-015-0289-2.

Sahu J, Sarmah R, Dehury B, Sarma K, Sahoo S, Sahu M,et al. 2012. Mining for SSRs and FDMs from expressed sequence tags of *Camellia sinensis*. Bioinformation 8:260-6. doi:10.6026/97320630008260.

Tzeng YM, Geethangili M. 2011. Review of pharmacological effects of A*ntrodia camphorata* and its bioactive compounds. Evidence-based Complement. Altern Med 2011. doi:10.1093/ecam/nep108.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3 —new capabilities and interfaces. Nucleic Acids Res 40:1-12. doi:10.1093/nar/gks596.

Victoria FC, da Maia LC, de Oliveira AC. 2011. In silico comparative analysis of SSR markers in plants. BMC Plant Biol 11:15. doi:10.1186/1471-2229-11-15.

Wu CC, Chu FH, Ho CK, Sung, CS, Chang SH. 2017. Comparative analysis of the complete chloroplast genomic sequence and chemical components of *Cinnamomum micranthum* and *Cinnamomum kanehirae*. Holzforschung 71: 189-197. doi:10.1515/hf-2016-0133.

Wu CC, Ho CK, Chang SH. 2015. Genomics study in Cinnamomum kanehirae Hayata. The Conference of Forestry Conservation and Sustainable Development. Taipei, Taiwan: Taiwan Forestry Research Institute.

Wu CC, Ho CK, Chang SH. 2016. The complete chloroplast genome of *Cinnamomum kanehirae* Hayata (Lauraceae). Mitochondrial DNA part A 27: 2681-82. doi:10.3109/19401736.2015.1043541.

Wu T, Luo S, Wang R, Zhong Y, Xu X, Lin Y, et al. 2014. The first Illumina-based *de novo* transcriptome sequencing and analysis of pumpkin (*Cucurbita moschata* Duch.) and SSR marker development. Mol Breed 34:1437-47. doi:10.1007/s11032-014-0128-x.

Xiao Y, Zhou L, Xia W, Mason AS, Yang Y, Ma Z, Peng M. 2014. Exploiting transcriptome data for the development and characterization of gene-based SSR markers related to cold tolerance in oil palm (*Elaeis guineensis*). BMC Plant Biol 14:384. doi:10.1186/s12870-014-0384-2.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821-9. doi:10.1101/gr.074492.107.

Zhang H, Wei L, Miao H, Zhang T, Wang C. 2012a. Development and validation of genic-SSR markers in sesame by RNA-seq. BMC Genom 13:316. doi:10.1186/1471-2164-13-316.

Zhang L, Zuo K, Zhang F, Cao Y, Wang J, Zhang Y, Sun X, Tang K. 2006. Conservation of noncoding microsatellites in plants: implication for gene regulation. BMC Genom 7:323. doi:10.1186/1471-2164-7-323.