Discovery of Simple Sequence Repeat Markers through Transcriptome Analysis of *Baccaurea motleyana*

Khairun Hisam Nasir¹, Muhammad Fairuz Mohd Yusof³, Mohd Shahril Firdaus Abdul Razak¹, Siti Norsaidah Ibrahim¹, Mira Farzana Mohamad Moktar¹, Khadijah Awang² and Mohd Shukri Mat Ali³

¹. Centre for Molecular Marker Discovery and Validation (CMDV), Agri-Omic dan Bioinformatics, Biotechnology and Nanotechnology Research Centre, Malaysian Agricultural Research and Development Institute (MARDI), Serdang 43400, Selangor, Malaysia
². Industrial Crop Research Centre, Malaysian Agricultural Research and Development Institute (MARDI), Serdang 43400, Selangor, Malaysia
³. Horticulture Research Centre, Malaysian Agricultural Research and Development Institute (MARDI), Serdang 43400, Selangor, Malaysia

Abstract: *Baccaurea motleyana* (rambai) is underutilized fruits that are native to Malaysia, Indonesia and Thailand. In this study, a total of 54,779 unigenes identified from rambai transcriptome were used for simple sequence repeat (SSR) analysis by MicroSatellite (MISA). A total of 20,420 SSRs were found to be distributed within 37.27% of the total number of unigenes. Mononucleotide repeats represented the main type, accounting for 64.04%, followed by trinucleotide repeats (20.28%) and dinucleotide repeats (19.94%). Gene annotation to seven databases has a success ratio of 68.53% (National Center for Biotechnology Information (NCBI) protein sequences), 53.68% (NCBI nucleotide sequences), 27.43% (Kyoto Encyclopedia of Genes and Genome Ortholog), 56.0% (SwissProt), 52.44% (Protein family), 53.99% (Gene Ontology) and 26.44% (Kluster of Orthologous Group). Further rambai SSRs were randomly selected and validated to *B. motleyana* (rambai), *B. macrocarpa* (tampoi), *B. polyneura* (jentik-jentik), *B. ramiflora* (pupor) and *B. scortechinii* (setambun).

Key words: *Baccaurea*, SSR, transcriptome, gene annotation and SSR validation.

1. Introduction

*Baccaurea motleyana* (rambai) is an indigenous fruit species of Malaysia. It belongs to the Phyllanthaceae family, and the five species, namely *B. macrocarpa* (tampoi), *B. polyneura* (jentik-jentik), *B. ramiflora* (pupor), *B. lanceolata* (asam pahong) and *B. angulata* (tampoi belimbing) are found in abundance in the forest in Peninsular Malaysia, whereas *B. angulata* (tampoi belimbing) are found in abundance in the forest in Peninsular Malaysia, whereas *B. angulata* (tampoi belimbing) are found in abundance in the forest in Peninsular Malaysia, whereas *B. angulata* (tampoi belimbing) are found in abundance in the forest in Peninsular Malaysia, whereas *B. angulata* (tampoi) is an endemic to Borneo which can be found wild and also cultivated [1]. Currently, rambai is widely cultivated in the orchard, home gardens, and many of these trees are still found in the forest. Rambai trees are mainly planted from seeds.

They are valued for their shade and fruits, which are eaten raw, stewed or made into jam or wine. The fruit skin dries under the sun for 2-3 days and can be used in dishes for seasoning, especially in Malay community.

An effort in collecting and conserving rambai has been carried out by Industrial Crop Research Centre, Malaysian Agriculture and Research Centre (MARDI) to determine the distribution, diversity and elite accessions. At present, there are 155 accessions of rambai that have been characterized, and 25 accessions have been planted at MARDI Station Jelebu, Negeri Sembilan. There are some variations among the accessions based on morphological characters, especially fruit characters. Normally, those of selected cultivated fruit trees are sweet, sweet sour and pleasant to eat.

Corresponding author: Khairun Hisam Nasir, Ph.D., research fields: molecular marker, gene expression, gene silencing, transient assay.
Rambai still lacks the required basic genomic resources for efficient breeding. Study of de novo transcriptome assembly of the arillode of *B. motleyana* has been carried out, and the assembled contigs of transcriptome are deposited in the DDBJ TSA [2]. To the best of our knowledge, no report on rambai SSR has been published to date.

Simple sequence repeat (SSR) is arrays of short motifs that can be characterized by their hyper variability, abundance, reproducibility, Mendelian inheritance and codominant nature [3, 4]. SSRs can be harvested from genomics and transcriptomes that are known as genomic SSRs and EST-SSRs. Since EST-SSRs derived from expressed sequences are more evolutionarily conserved compared to non-coding sequences, this type of SSR has a relatively high transferability than genomic SSRs [5]. Development of next-generation sequencing technologies assists researchers to harvest and evaluate many EST-SSRs from important crops such as *Ipomoae batatas* [6], *Cicer arietinum* [7], *Epimedium sagittatum* [8], and *Elymus sibiricus* [9].

In this study, transcriptome data obtained from Illumina paired end sequencing were used to mine rambai EST-SSRs and further validated to five *Baccaurea* species: *B. motleyana* (rambai), *B. macrocarpa* (tampoi), *B. polyneura* (jentik-jentik), *B. ramiflora* (pupor) and *B. scortechinii* (setambun).

### 2. Material and Methods

#### 2.1 Plant Material for RNA

Sample used for RNA extraction in this study was grown at Industrial Crop Research Centre, Malaysian Agriculture Research and Development Institute (MARDI), Selangor, Malaysia. The *B. motleyana* (rambai) apical meristems were collected and frozen in liquid nitrogen until further use.

#### 2.2 RNA Extraction

The rambai shoot apical meristems were ground into fine powder with pestle and mortar that had been pre-cooled with liquid nitrogen. One milliliter of CTAB extraction buffer was added into ~100 mg tissue and incubated at 65 °C for 30 min before being centrifuged at 16,000×*g* for 20 min. The supernatant collected was used for chloroform:isoamyl (24:1) extraction with equal volume (1 volume supernatant:1 volume chloroform:isoamyl (24:1). In total 8 M LiCl (1/3 volume of supernatant), was added into the supernatant and incubated overnight at 4 °C. The solution was centrifuged at 16,000×*g* for 60 min at 4 °C. Pellets were washed with 70% ethanol and centrifuged for 15 min before being dried at room temperature. The pellets were resuspended in 20 μL nuclease-free water and treated with DNase I Recombinant, RNase-free water (QIAGEN, USA) to prevent the sample from DNA contamination. The quality and quantity of extracted RNA were examined using a 2100 Bioanalyzer (Agilent Technology, Santa Clara, CA, USA). The RNA sample with RNA integrity number > 7 was used for Illumina sequencing.

#### 2.3 Plant Material for DNA Extraction

Five *Baccaurea* species, *B. motleyana* (rambai), *B. macrocarpa* (tampoi), *B. polyneura* (jentik-jentik), *B. ramiflora* (pupor) and *B. scortechinii* (setambun) were collected at MARDI Station Jelebu, Negeri Sembilan and used for DNA extraction. *Baccaurea* leaves were DNA-extracted using a conventional plant DNA extraction method. Approximately 5 g of leaves were stuffed into a 2 mL tube with stainless steel balls and ground to a powder using Tissue Lyser (Qiagen, GmbH Germany). Next, 2,000 μL of the extraction buffer (final concentration: polyvinylpyrrolidone (PVP) 2%, diethyldithiocarbamate (DIECA) 4 mM, 5 mM ascorbic acid, 1.4 M NaCl, 100 mM tris-HCl (pH 8.0) and 20 mM ethylenedinitrilotetraacetic acid (EDTA)) was added to the ground powder. The mixture was incubated at 65 °C for 1 h with intermittent mixing. The same volume of isopropanol was added to this mixture, and the two were
homogenized by inverting the plates. The mixture was centrifuged at 5,500 rpm for 15 min. Later the supernatant was transferred into an equal volume of isopropanol and centrifuged at 5,500 rpm for 15 min. The DNA was air-dried and resuspended in 50 μL tris-EDTA buffer; the DNA was quantified for concentration using Thermo Labsystems Fluoroskan Ascent™ (Thermo Scientific, USA); and DNA integrity was tested on 0.8% agarose gel for quality control.

2.4 PCR SSR Validation

Primers were successfully designed and oligo synthesized for 28 SSRs after SSR mining; these primers were used for PCR. The M13-tailed primer was used, which includes three primers: a forward primer (with an M13 tail sequence), a reverse primer, and a fluorescently labeled M13 IRDye primer [10]. PCR amplification was performed in a total reaction volume of 10 μL, consisting of 1.0 μL of template DNA (approximately 40 ng·µL−1), 1 μL PCR reaction buffer (1.5-3.0 mM MgCl2), 2 mM dNTP mixture, 10 μmol of each primer set, 5 μmol of a fluorescent dye (FAM/VIC/NED and PET), and 1 unit of Taq DNA polymerase (Invitrogen, USA). Amplification was performed using a Peltier Thermal Cycler, DNA Engine Tetrad 2 (BioRad, USA). The PCR profile was an initial denaturation at 94 °C for 5 min, followed by 34 cycles of 1 s denaturation at 94 °C for 30 s, annealing (see Table 7) for 45 s, and 72 °C for 45 s extensions and finally 1 cycle of 72 °C for 7 min. The PCR products were genotyped using the 3730xl DNA Analyzer (Applied Biosystems, USA).

2.5 Sequence Data Analysis and De Novo Transcriptome Assembly

The original raw data from Illumina HiSeq™ were filtered first to generate high-quality data. The filtering included removal of adaptor contaminants, removal of reads with more than 10% of N, where N represented the base with uncertainty, and removal of low-quality base reads, with the quality value Q ≤ 20. The high-quality transcriptome data were then assembled using Trinity [11] version r20140413p1 with parameter minimum k-mer coverage of two and others being default. De novo assembled transcripts from Trinity were clustered using Corset Version 1.05 to remove redundancy.

2.6 Molecular Marker Discovery

SSR markers were discovered using MicroSAtellite (MISA), an identification tool with the following parameters: minimum SSR motif length of 12 base pairs (bp) and repeat lengths of mono-10, di-6, tri-5, tetra-5, penta-5 and hexa-5. The maximum size of interruption allowed between two different SSRs in a compound was 0 bp. Identified SSR markers were then mined using a custom Perl script and R programming to examine the repeat distributions. Mining of SSR markers included choosing SSR markers that represented more than two alleles and represented a single contig. We extracted 200 bp of each side of repeat motif region using BEDTools Version 2.21.0. SSR markers were annotated using the BLAST program (blastx) against a non-redundant (nr) protein database in NCBI with an E-value of $1 \times 10^{-5}$ and BLAST2GO software (Supplementary excel file of rambai SSR).

2.7 Gene Functional Annotation

The functional annotation was outsourced to Novogen Company. The software and parameters used in each database were, NT using NCBI Basic Local Alignment Search Tool (BLAST version 2.2.28), the E-value threshold of 1.0e-5, non-redundant protein sequences (NR) using SwissProt and Cluster of Orthologous Groups (KOG) of Diamond 0.8.22. NR and SwissProt databases, the E-value thresholds are 1.0e-5 and 1.0e-3 for KOG. Protein family (Pfam), the prediction of protein structure domain was made using HMMER 3.0 package, hmmscan, the E-value threshold is 0.01.
Gene ontology (GO) for the protein annotation results of NR and Pfam using Blast2GO v2.5 [12] and a Perl script, the E-value threshold is 1.0e-6 and Kyoto Encyclopedia of Genes and Genome (KEGG) using automatic annotation server, the E-value threshold is 1.0e-10.

2.8 Scoring and Data Analyses

The fragment files were analyzed using GeneMapper 5.0 (Applied Biosystems). GS500LIZ was used as the standard control. Allele peaks in the electropherograms were scored and analyzed [13]. The scored peaks were imported into a Microsoft Excel file, and the data were pre-analyzed using MicroChecker v2.2.3 [14] to check for the presence of null alleles, large allele dropouts, and scoring errors due to stutter peaks. The passed data, after MicroChecker analysis, were converted into an input data file using CONVERT 1.2 [13] for further analysis. POPGENE32 version 1.31 was used to calculate the number of alleles, allele frequencies of each variety, and the genetic distance between varieties. NTSYSpc software was used to construct a dendrogram using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) method based on the genetic distance produced from the POPGENE32 software.

3. Result and Discussion

3.1 Sequencing and De Novo Assembly of Rambai Transcriptome

High-throughput next-generation sequencing generated a total of 81,941,472 paired end raw reads of rambai transcriptome comprising 12.3 Gigabase (Gb) with an average read length of 150 base pair (bp). After quality assessment, approximately 96.5% of high-quality data were trimmed out based on the base quality and adaptor contaminants (Table 1). High-quality cleaned reads (79,121,624 paired end reads) were then assembled into 54,813 transcripts and 54,779 unigenes in the rambai transcriptome (Table 2).

| Type of data | Total reads | Total bases (base pair) | Average read length (base pair) |
|--------------|-------------|-------------------------|-------------------------------|
| Raw data     | 81,941,472  | 12.3G                   | 150                           |
| Clean data   | 79,121,624  | 11.9G                   | 150                           |
| Percentage of high quality data (%) | 96.5 | 96.7 | 100 |

Table 1  Statistics of rambai sequence data quality control.

| Type of data | Total reads | Total bases (base pair) | Average read length (base pair) |
|--------------|-------------|-------------------------|-------------------------------|
| Transcripts  | 54,813      | 54,779                  |                               |
| Unigenes     | 61,340,536  | 61,331,921              |                               |
| Range of length transcripts/unigenes (base pair) | 201-9,101 | 201-9,101       |
| Average of length transcripts/unigenes (base pair) | 1,119 | 1,120 |
| Length of N50 transcripts/unigenes (base pair) | 1,537 | 1,538 |

Table 2  Statistics of de novo transcriptome assembly.

| Type of data | Transcripts a | Unigenes b |
|--------------|---------------|------------|
| Number of transcripts/unigenes | 54,813 | 54,779 |
| Total bases transcripts/unigenes (base pair) | 61,340,536 | 61,331,921 |
| Range of length transcripts/unigenes (base pair) | 201-9,101 | 201-9,101 |
| Average of length transcripts/unigenes (base pair) | 1,119 | 1,120 |
| Length of N50 transcripts/unigenes (base pair) | 1,537 | 1,538 |

a Transcripts were the transcripts produced by transcriptome assembler Trinity.  
b Unigenes were the clustering transcriptomes that were not redundant.  
c The N50 length was defined as the length for which the collection of all transcripts of that length or longer contains at least half of the total of the lengths of the transcripts and for which the collection of all transcripts of that length or shorter contains at least half of the total of the lengths of the transcripts.
3.2 Molecular Marker Discovery

A total of 20,420 of SSR markers (Table 4) were distributed in the rambai transcriptome accounting for 37.28% of the total unigenes. The total length of SSR markers discovered in the rambai transcriptome was 61,331,921 bp (Table 3). All SSR markers were divided into mono-, di-, tri-, tetra-, penta- and hexa-nucleotide repeats. Mononucleotide repeats were the most abundant (64.05%), followed by trinucleotide repeats (20.28%) and dinucleotide repeats (19.94%) (Table 4). The mononucleotide repeats and compound formation were excluded from the analysis. The most frequent repeat motifs for trinucleotide were GAA/TTC, followed by dinucleotide repeats AG/CT and tetranucleotide repeats AAAT/ATT (Table 4).

3.3 Gene Functional Annotation

To get comprehensive gene function of rambai genes, gene annotation to seven databases, namely National Center for Biotechnology Information (NCBI) non-redundant protein sequences (NR), NCBI nucleotide sequences (NT), Kyoto Encyclopedia of Genes and Genome Ortholog (KO), SwissProt, Protein family (Pfam), Gene Ontology (GO) and Cluster of Orthologous Groups (KOG), was used as reference. The NCBI non-redundant protein sequences (NR), include protein sequence information from GenBank, Protein Data Bank (PDB), SwissProt, Protein Information Resource (PIR) and Protein Research Foundation (PRF). The NCBI nucleotide sequences (NT) are the nucleotide sequence database that includes nucleotide sequence from GenBank of the European Bioinformatics Institute (EMBL) and DNA Data Bank of Japan (DDBJ). KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as cell, organism and ecosystem, from molecular-level information, especially for large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. KEGG is an established Cluster of Orthologous (KO) annotation system that can accomplish the function annotation of the genome/transcriptome of a newly sequenced species. SwissProt is a manual annotated and reviewed protein sequence database that has a high-quality protein sequence database from experimental results, computed features and scientific conclusions. Pfam is comprehensive collection of protein domains and families, represented as multiple sequence alignments and as profile of hidden Markov models. Many proteins are composed of structural domains, and the protein sequence of a specific

Table 3  Summary of rambai SSR mining from unigene.

| Total number of sequences examined | 54779 |
| Total size of examined sequences (bp) | 61,331,921 |
| Total number of identified SSRs | 20,420 |
| Number of SSR containing sequences | 14,883 |
| Number of sequences containing more than 1 SSR | 4,069 |
| Number of SSRs present in compound formation | 1,942 |

Table 4  Summary of frequent repeat motif rambai SSR markers.

| Type of SSR markers | Total number | Total length (base pair) | Most frequent repeat motif |
|---------------------|--------------|--------------------------|---------------------------|
| SSR markers type p1 | 13,079       | 133,708                  | A/T                       |
| SSR markers type p2 | 2,968        | 29,864                   | AG/CT                     |
| SSR markers type p3 | 4,141        | 46,536                   | GAA/TTC                   |
| SSR markers type p4 | 202          | 2,472                    | AAAT/ATT                  |
| SSR markers type p5 | 15           | 290                      | ATAAC/GTTAT               |
| SSR markers type p6 | 15           | 390                      | ATGGTG/ACCATC             |
| Total              | 20,420       | 213,260                  |                           |
structural domain possesses a certain degree of conservative property. GO is the established standard for the functional annotation of gene products and controlled vocabulary used to classify the functional attributes of gene products of a biological process, a molecular function and a cellular component.

The success ratios of rambai gene annotation were 68.53% (NR), 53.68% (NT), 27.43% (KO), 56.00% (SwissProt), 52.44 (Pfam), 53.99% (GO) and 26.44% (KOG) (Table 5). The genes that were successfully annotated in seven databases were 13.77%, and only 75.84% were successfully annotated in one database (Table 5).

Gene annotation to NR showed higher similarity to *Jatropha curcas* (26.4%), followed by *Ricinus communis* (15.3%), *Populus trichocarpa* (14.3%), *Populus eupharatica* (11.8%), *Theobroma cacao* (4.3%) and others (27.8%) (Fig. 1).

Gene annotation to GO has grouped rambai genes into three main GO domains: biological process, cellular component and molecular function. In the biological process, more than 75,000 genes were identified, and three highest gene functions were involved in the cellular process, the metabolic process and the single-organism process; in the cellular component, more than 46,000 genes were identified, and three highest gene functions were involved in cell, cell part and organelle; in the molecular function, more than 10,000 genes were identified, and three highest gene functions were involved in binding, catalytic activity and transport activity (Fig. 2).

The Cluster of Orthologous Groups (KOG) to rambai genes has been classified into 26 KOG groups (Fig. 3). The three highest groups with more than 9% of genes annotated were attributed to general function prediction, posttranslational modification/protein turnover/chaperones and translation/ribosomal structure/biogenesis.

KEGG has annotated rambai genes into five KEEG pathways: the cellular process, environmental processing, genetic information processing, metabolism, and the organismal system (Fig. 4). The highest genes were in metabolism (6,741 genes), followed by genetic information processing (3,247 genes), the cellular process (706 genes), the organismal system (550 genes) and environmental information processing (458 genes).

Genes containing SSRs were annotated into the Clusters of Orthologous (KO) (Fig. 5). In the cellular process, four major annotate genes were involved in endocytosis, peroxisome, phagosome and regulation of autophagy; in environmental information processing, three major annotated genes were involved in plant hormone signal transduction, the phosphatidylinositol signal system and ABC transports; in genetic information processing, three highest annotated genes were involved in protein processing in the endoplasmic reticulum, ribosome and spliceosome; in
Discovery of Simple Sequence Repeat Markers through Transcriptome Analysis of *Baccaurea motleyana*

Fig. 2 Gene function classification (GO) of rambai.
Discovery of Simple Sequence Repeat Markers through Transcriptome Analysis of *Baccaurea motleyana*

Fig. 3  The classifications of genes annotated in KOG. The x-axis is the 26 KOG group, and y-axis is the percentage of genes annotated under this group in the total annotated genes.

Fig. 4  Kyoto Encyclopedia of Genes and Genome (KEGG) classification. The y-axis is the names of KEGG pathways, and x-axis is the number of the genes annotated in the pathway and the ratio between the number in this pathway and the total number of annotated genes. The KEGG metabolic pathways gene involved in function annotation and divided into five branches: A—cellular processes, B—environmental information processing, C—genetic information processing, D—metabolism, and E—organisal systems.
Discovery of Simple Sequence Repeat Markers through Transcriptome Analysis of Baccaya motleyana

Fig. 5 Cluster of orthologous (KO) annotation of gene-containing SSRs.
Table 5  The success ratio of annotated genes.

| No. | Database annotation                                         | Number of unigenes | Percentage (%) |
|-----|------------------------------------------------------------|--------------------|----------------|
| 1   | NCBI non-redundant protein sequences (NR)                  | 37,541             | 68.53          |
| 2   | NCBI nucleotide sequences (NT)                             | 29,406             | 53.68          |
| 3   | KEGGortholog (KO)                                          | 15,027             | 27.43          |
| 4   | SwissProt                                                  | 30,678             | 56.00          |
| 5   | Protein family (Pfam)                                      | 28,728             | 52.44          |
| 6   | Gene ontology (GO)                                         | 29,576             | 53.99          |
| 7   | Cluster of orthologous groups (KOG)                        | 14485              | 26.44          |
| 8   | The number of genes successfully annotated in all the seven databases | 7,545              | 13.77          |
| 9   | The number of genes successfully annotated in at least one database | 41,547             | 75.84          |
| 10  | Total unigenes                                             | 54,779             | 100            |

Table 6  Prediction pathway levels of gene-containing SSRs from rambai. The highest pathway level is from the cellular process, environment information processing, genetic information processing, metabolism and the organelle system with ten selected genes.

| 1st-level (pathway) | 3rd-level (pathway) | Description of gene-containing SSRs                                                                 |
|---------------------|---------------------|-----------------------------------------------------------------------------------------------------|
| Cellular processes  | Endocytosis         | 1-phosphatidylinositol-4-phosphate 5-kinase, phospholipase D1/2, dynamin GTPase, heat shock 70kDa protein 1/8, clathrin heavy chain, actin related protein 2/3 complex, subunit 5, ADP-ribosylation factor 1, capping protein (actin filament) muscle Z-line, alpha, capping protein (actin filament) muscle Z-line, beta |
| Environmental       | Plant hormone signal transduction | phytochrome-interacting factor 3, protein brassinosteroid insensitive 1, transcription factor MYC2, pathogenesis-related protein 1, auxin influx carrier (AUX1 LAX family), transcription factor TGA, ABA responsive element binding factor, auxin responsive protein IAA, auxin response factor, auxin responsive GH3 gene family |
| Genetic information processing | Protein processing in the endoplasmic reticulum | mannosyl-oligosaccharide glucosidase, peptide-N4-(N-acetyl-beta-glucosaminyl) asparagine amidase, S-phase kinase-associated protein 1, translation initiation factor 2 subunit 1, cullin 1, RING-box protein 1, molecular chaperone HtpG, ubiquilin, ubiquitin-conjugating enzyme, dolichyl-diphosphooligosaccharide--protein glycosyltransferase, malate dehydrogenase, isocitrate dehydrogenase (NAD+), 6-phosphogluconate dehydrogenase |
metabolism, three highest annotated genes were involved in carbon metabolism, biosynthesis of amino acids and starch/sucrose metabolism; in the organismal system, two major genes were involved in circadian rhythm and plant-pathogen interaction.

Further, we made analysis to predict what genes were involved from the first level pathway of the cellular process, environment information processing, genetic information processing, metabolism, and the organismal system and the third level pathway of endocytosis, plant hormone signal transduction, protein processing in the endoplasmic reticulum, carbon metabolism and plant-pathogen interaction. A list of ten selected genes is provided in Table 6.

3.4 Rambai SSR Validation

Twenty-eight rambai SSRs were randomly selected and used for validation of five Baccarea species, namely B. motleyana (rambai), B. macrocarpa (tampoi), B. polynera (jentik-jentik), B. ramiflora (pupor) and B. corotechnii (setambun). Of 28 SSRs, only four were polymorphic, and the call rate was more than 90% across Baccarea species. The remaining microsatellites were monomorphic, exhibited low call rates, and produced multiple bands across the rambai species. Analysis of four microsatellites revealed a total of 52 alleles, which ranged from 8 (B9) to 21 (B89) alleles per locus and had a mean of 13.75 alleles per locus. The range of the polymorphic information content (PIC) value was from 0.5656 (B9) to 0.9402 (B89), with an average of 0.8087, whereas the value of gene diversity and heterozygosity ranged from 0.5851 (B89) to 0.9432 (B89) and from 0.1667 (B89) to 0.4545 (B89), respectively (Table 8).

Dendogram analysis of 15 accessions from five Baccarea species was classified into two main groups (Fig. 6). Group I can be classified into two subgroups. Subgroup Ia consists of Tampoi kersa MDI 1, Tampoi burung, Tampoi MDIT 37, Tampoi MDIT 35, Tampoi MDIT 31, Rambai Langkawi MDIRL 42, Tampoi MDIT 11 and MDIT 6, and subgroup 1b consists of only Tampoi MDIT 5. Group II consists two subgroups, with subgroup IIa consisting only Setambun and subgroups IIb consisting of Tampoi tungau MDITG3, MDIT 2, Rambai Sarawak MDIR 21, jentik-jentik JJ1 and pupor MDIP.
| No. | Seq ID       | Forward                  | Reverse                  | Annealing temperature (°C) | SSR repeat | Product size (bp) |
|-----|-------------|--------------------------|--------------------------|----------------------------|------------|-------------------|
| 1   | B1_Cluster-10174.0 | 5'-TGAATCCACGGTTTGTT-3'  | 5'-TTTGGTTGTTGTGATCTTG-3' | 55.0                       | (TC)6      | 154               |
| 2   | B4_Cluster-10604.0 | 5'-GGCTCTCTCTCTTCTTAAC-3' | 5'-GTGTTGCTTGGTTGCTTTA-3' | 55.0                       | (TC)7      | 137               |
| 3   | B6_Cluster-10822.0 | 5'-TTCTTCCACGAATTTATCAC-3' | 5'-TCACTGTCGATCACATTCACT-3' | 55.0                       | (TC)7      | 157               |
| 4   | B7_Cluster-11256.0 | 5'-GTAGCAGACTATCGAATGG-3' | 5'-GTAATCCACGGTTTGTTGATCTTG-3' | 55.0 | (AT)6 | 143 |
| 5   | B10_Cluster-11505.0 | 5'-TTGTGCTGATACCACTGTCA-3' | 5'-GTACACGCTGACACCAATCC-3' | 55.0                       | (AG)6      | 172               |
| 6   | B11_Cluster-11521.1 | 5'-TCCTACATAGCAACACACAG-3' | 5'-ACTCTTTTCTCCAGAATTTCT-3' | 55.0                       | (AT)7      | 132               |
| 7   | B13_Cluster-11549.0 | 5'-TTTATTAGCACTGAGCAAGA-3' | 5'-CCCTTCGATGTGTAATTTCT-3' | 55.0                       | (AC)8      | 148               |
| 8   | B19_Cluster-11839.1 | 5'-GTGATAACAGTGCTGACAT-3' | 5'-GTACACGCTGACACCAATCC-3' | 55.0                       | (TC)6      | 149               |
| 9   | B29_Cluster-12306.0 | 5'-GTTCATAGGGCAATTTGA-3'  | 5'-AGGAGAACGGATATCTGAGAG-3' | 55.0                       | (AG)6      | 149               |
| 10  | B32_Cluster-1261.0  | 5'-GTAATGACATACGCTTCAA-3' | 5'-ACATTATCCCATATTGACTC-3' | 55.0                       | (TG)8      | 153               |
| 11  | B34_Cluster-13088.0 | 5'-AAAAGGGTGGGATGTCAA-3'  | 5'-GCTTCGGCTCCTTTGTTTCT-3' | 55.0                       | (AG)6      | 145               |
| 12  | B37_Cluster-13557.0 | 5'-AACTGTTGCTGGGATGAC-3'  | 5'-CCTCTTACAACTTACCTGTC-3' | 55.0                       | (CT)6      | 157               |
| 13  | B38_Cluster-13622.1 | 5'-CACTGTTGCTCCTCTTAAGA-3' | 5'-AAAACGATGAGTTGGGAGAG-3' | 55.0                       | (TC)6      | 129               |
| 14  | B39_Cluster-13704.0 | 5'-TCTGTGTTGCAAGATCAATG-3' | 5'-GGAATTGCTTTGTGATCTT-3' | 55.0                       | (TC)6      | 150               |
| 15  | B45_Cluster-14258.0 | 5'-TTTGTTGTGTTGTGTTGTT-3' | 5'-CCCTTAATGGAACCCCAAAC-3' | 55.0                       | (TG)8      | 122               |
| 16  | B55_Cluster-10294.0 | 5'-TCAAAACCACATCTTCTTCT-3' | 5'-AATCATCTTCTTTGATGAC-3' | 55.0                       | (AGC)6     | 125               |
| 17  | B59_Cluster-10415.0 | 5'-TCGGTTACATGTCAAACAC-3' | 5'-GTAATGGGACATACATTTTCT-3' | 55.0                       | (TTA)5     | 156               |
| 18  | B61_Cluster-10592.0 | 5'-ACAAGAGTTGCTCAGAAAGA-3' | 5'-TAGTTTGCTGGCACACCAAGT-3' | 55.0                       | (TGG)5     | 153               |
| 19  | B66_Cluster-10721.1 | 5'-TTTCACTGACATCTTCTAC-3' | 5'-ACTGAGCCATATGACACAC-3' | 55.0                       | (AAT)5     | 150               |
| 20  | B69_Cluster-108.0   | 5'-GCTCTTCCACTCTTTGG-3'   | 5'-GACACACACCTTTCAATG-3'   | 55.0                       | (CAA)5     | 133               |
| 21  | B79_Cluster-1414.0  | 5'-GAAAGCTTCACTCGACAGC-3' | 5'-CCAATGAGCTTCTGCAGTTG-3' | 55.0                       | (ATCA)5    | 166               |
| 22  | B84_Cluster-1396.0  | 5'-CTGGATGTCGATGATGATA-3' | 5'-CTGCAATGATCTCTTGTTA-3' | 55.0                       | (TGTGTT)4  | 162               |
| 23  | B87_Cluster-16566.0 | 5'-CTCTGTTGCTTTGTTGTTG-3' | 5'-CTGATCTCTCTCCCACAGAAGA-3' | 55.0 | (TGGTG)4 | 146 |
| 24  | B89_Cluster-16871.0 | 5'-ACACAGGACACTCCCTTTTT-3' | 5'-CAAATCTAATCATGACACAT-3' | 55.0                       | (AAAAC)4   | 166               |
| 25  | B90_Cluster-17149.0 | 5'-TACACGAGATTGTGCTTTG-3' | 5'-GGTGCTCTCTTCTCCATACA-3' | 55.0                       | (AGA)6     | 161               |
| 26  | B98_Cluster-49.10254 | 5'-CACTCCTCACGTCATCATAC-3' | 5'-CATATGCTGGGAATGACAC-3' | 55.0                       | (CCCGTG)5  | 151               |
| 27  | B99_Cluster-49.11456 | 5'-GCTCTCCTTTCTCTTTTCT-3' | 5'-GCCAGTCAAGAGCTCCCAA-3' | 55.0                       | (AACAAG)4  | 139               |
| 28  | B100_Cluster-49.12264| 5'-GGAGATTTAAAAACAGGACTGAC-3' | 5'-GCTTTAAAGAATCCCACATC-3' | 55.0                       | (AG)8      | 137               |
Table 8  Characterization of polymorphic rambai SSR.

| Marker | Major allele frequency | Allele No. | Availability | Gene diversity | Heterozygosity | PIC       |
|--------|------------------------|------------|--------------|----------------|----------------|-----------|
| B9     | 0.6250                 | 8.0000     | 0.8276       | 0.5851         | 0.1667         | 0.5656    |
| B17    | 0.2857                 | 9.0000     | 0.7241       | 0.8333         | 0.1905         | 0.8142    |
| B21    | 0.1364                 | 17.0000    | 0.7586       | 0.9205         | 0.2273         | 0.9149    |
| B89    | 0.0909                 | 21.0000    | 0.7586       | 0.9432         | 0.4545         | 0.9402    |
| Mean   | 0.2845                 | 13.7500    | 0.7672       | 0.8205         | 0.2597         | 0.8087    |

Fig. 6  Dendogram of five Baccaurea species: B. motleyana (rambai), B. macrocarpa (tampoi), B. polyneura (jentik-jentik), B. ramiflora (pupor) and B. scortechinii (setambun).

4. Conclusion

We have mined > 20,000 SSRs from a rambai leaf transcriptome. The trinucleotide represents the highest repeats, followed by the dinucleotide, the tetranucleotide and the pentanucleotide. Gene annotation to seven databases showed success ratios of 68.53% (NCBI protein sequences), 53.68% (NCBI nucleotide sequences), 27.43% (Kyoto Encyclopedia of Genes and Genome Ortholog), 56.0% (SwissProt), 52.44% (Protein Family), 53.99% (Gene Ontology) and 26.44% (Kluster of Orthologous Group). Validation of rambai SSRs to B. motleyana (rambai), B. macrocarpa (tampoi), B. polyneura (jentik-jentik), B. ramiflora (pupor) and B. Scortechinii (setambun) has group into two main groups.

Acknowledgment

The project was funded by the development fund of eleven Malaysian plans through Management and Germplasm Conservation of Genetic Resources Program Centre and Seed Gene Bank of the Malaysian Agricultural Research and Development Institute, 43400 Serdang, Selangor, Malaysia.

Author Contributions

Khairun Hisam Nasir collected rambai leaves, scored SSR fragments and wrote the paper; Muhammad Fairuz Mohd Yusof and Mira Farhana conducted bioinformatic analysis; and Mohd Shahril Firdaus analyzed the data. Siti Norsaidah performed the SSR experiment; Khadijah Awang provided the
rambai tree; and Mohd Shukri Mat Ali was the project leader.

References

[1] Khadijah, A., Mirfat, A. H. S., Mohd Nor, A., Siti Sofiah, M., Mohd Khairuddin, O., and Mohd Shukri, M. A. 2018. “Tropical Fruits ‘The Next Golden Crop for Sabah’.” Department of Agriculture Sabah Seminar.

[2] Deden, D. M., Arya, W. R., Azis, N., Roedhy, P. S., Winarso, D. W., and Eiichi, I. 2019. “Dataset of the First De Novo Transcriptome Assembly of the Arillode of Baccaurea motleyana.” Science Direct Data in Brief 22: 332-5.

[3] Scott, K. D., Eggler, P., Seaton, G., Rossetto, M., Ablett, E. M., Lee, L. S., and Henry, R. J. 2000. “Analysis of SSRs Derived from Grape ESTs.” Theor. Appl. Genet. 100: 723-6.

[4] Gupta, P. K., Balyan, I. S., Sharma, P. C., and Ramesh, B. 1996. “Microsatellites in Plants: A New Class of Molecular Markers.” Curr. Sci. 70: 45-54.

[5] Wei, W., Qi, X., Wang, L., Zhang, Y., Hua, W., Li, D., Lv, H., and Zhang, X. 2011. “Characterization of the Sesame (Sesamum indicum L.) Global Transcriptome Using Illumina Paired-End Sequencing and Development of EST-SSR Markers.” BMC Genomics 12: 451-63.

[6] Wang, Z., Fang, B., Chen, J., Zhang, X., Luo, Z., Huang, L., Chen, X., and Li, Y. 2010. “De Novo Assembly and Characterization of Root Transcriptome Using Paired-End Sequencing and Development of cSSR Markers in Sweet Potato (Ipomoea batatas).” BMC Genomics 11: 726-39.

[7] Garg, R., Patel, R. K., Tyagi, A. K., and Jain, M. 2011. “De Novo Assembly of Chickpea Transcriptome Using Short Reads for Gene Discovery and Marker Identification.” DNA Res. 18: 53-63.

[8] Zeng, S., Xiao, G., Guo, J., Fei, Z., Xu, Y., Roe, B. A., and Wang, Y. 2010. “Development of an EST Dataset and Characterization of EST-SSRs in a Traditional Chinese Medicinal Plant, Epimedium sagittatum (Sieb. Et Zucc.) Maxim.” BMC Genomics 11: 94-104.

[9] Zhou, Q., Luo, D., Ma, L., Xie, W., Wang, Y., Wang, Y., and Liu, Z. 2016. “Development and Cross-Species Transferability of EST-SSR Markers in Siberian Wildrye (Elymus sibiricus L.) Using Illumina Sequencing.” Sci. Rep. 6: 266-78.

[10] Arif, I. A., Khan, H. A., Shobrak, M., Al Homaidan, A. A., Al Sadoon, M., Al Farhan, A. H., and Bahkali, A. H. 2010. “Interpretation of Electrophoretograms of Seven Microsatellite Loci to Determine the Genetic Diversity of the Arabian Oryx.” Genetics and Molecular Research 9 (1): 259-65. doi: 10.4238/vol9-1gmr714.

[11] Grabherr, M. G., Haas, B. J., Yassour, M., et al. 2011. “Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome.” Nature Biotechnology 29:644-52.

[12] Stefan, G., Garcia-Gómez, J. M., Javier, T., Tim, D. W., Shivashankar, H. N., Robles, M., Manual, T., Joaquín, D., and Ana, C. 2008. “High-Throughput Functional Annotation and Data Mining with the Blast 2GO Suit.” Nucleic Acids Research 36: 3420-35. https://doi.org/10.1093/nar/gkn176.

[13] Glaubitz, J. C. 2004. “Convert: A User-Friendly Program to Reformat Diploid Genotypic Data for Commonly Used Population Genetic Software Packages.” Molecular Ecology Notes 4 (2): 309-10.

[14] Van Oosterho, C., Hutchinson, W. F., and Wills, D. P. 2004. “Micro-checker: Software for Identifying and Correcting Genotyping Errors in Microsatellite Data.” Molecular Ecology Notes 4 (3): 535-8. https://doi.org/10.1111/j.1471-8286.2004.00684.x.