Journal of Tropical Forest Science 38(1): 104–117 (2026)
https://doi.org/10.26525/jtfs2026.38.1.104
ISSN: 0128-1283, eISSN: 2521-9847

MULTIPLEXED INTER-SIMPLE SEQUENCE REPEAT GENOTYPING BY SEQUENCING (MIG-SEQ) OF NUT PRODUCING *RUBROSHOREA* AND *SHOREA* SPECIES

Wong SY[1,*] & Ogary KM[1,2]

[1]*Institute of Biodiversity and Environmental Conservation, Universiti Malaysia Sarawak, Kota Samarahan, Sarawak, Malaysia*
[2]*Sarawak Forestry Corporation, Lot 218, KCLD, Jalan Tapang, Kota Sentosa, 93250 Kuching, Sarawak, Malaysia*

**sywong@unimas.my*

Submitted June 2025; accepted August 2025

Engkabang nuts from two Dipterocarpaceae genera, *Rubroshorea* and *Shorea*, are collected from semi-managed or wild trees in Borneo to produce oil (tengkawang and terendek oils) and vegetable fat (illipe butter) for own consumption or to be sold locally. Although the nuts of these species display similar fatty acid profiles, the species exhibit diverse morphological and physiological features which are influenced by DNA polymorphisms responsible for the differences in traits among these species. In this study, we determined genome-wide single-nucleotide polymorphisms (SNPs) among geographically diverse Engkabang accessions using the Multiplexed Inter-Simple Sequence Repeat Genotyping by Sequencing (MIG-seq) method. A total of 1,160 SNPs was identified in a collection of ninety-six Engkabang accessions of *Rubroshorea macrophylla*, *R. splendida*, and *Shorea seminis* from Sarawak. Genetic diversity analyses, based on the common SNP sites, divided *R. macrophylla* accessions into two main groups, Central Sarawak and West Sarawak (*R. splendida* and *S. seminis* embedded within). Single-nucleotide polymorphisms data generated in this work will facilitate further genetic studies on Engkabang breeding and production of DNA markers.

Keywords: genetic similarity, inter-simple sequence repeat (ISSR), Sarawak, single-nucleotide polymorphisms (SNPs), Tengkawang

INTRODUCTION

Trading of Engkabang (known as Tengkawang in Kalimantan, Indonesia) oil and fat in Borneo started in the 1800s, mainly for use in the manufacture of candles and as lubricating agent in steam machinery and tram cars. The vegetable fats from Engkabang are also used as a substitute for cocoa butter, and are known as Cocoa Butter Equivalents (CBE) (Blicher-Mathiesen 1994). Engkabang oil is produced from the nuts of *Shorea* species known as Borneo Illipe nuts (Smythies 1958). Recently, the genus *Shorea* was split into seven genera of which *Rubroshorea* (Meijer) P.S.Ashton & J.Heck. (Ashton & Heckenhauer 2022) with seventy-one species includes all the oil producing species, *viz. R. macrophylla* (de Vriese) P.S.Ashton & J.Heck., *R. mecistopteryx* (Ridl.) P.S.Ashton & J.Heck., *R. pinanga* (Scheff.) P.S.Ashton & J.Heck., *R. stenoptera* (Burck) P.S.Ashton & J.Heck., and *R. splendida* (de Vriese) P.S.Ashton & J.Heck. Of these, *R. stenoptera* is the most utilised since it has the largest fruit size, while *R. macrophylla* and *R. pinanga* are harvested because of their wide distribution (Fambayun & Hut 2014, Coolen 2013, Yulita 2016, Darmawan et al. 2020, Heri et al. 2020). Within *Shorea* in the new strict sense, *S. seminis* (de Vriese) Slooten and *S. sumatrana* Sym. ex Desch are also utilised for nut collection along with the lesser known species, *S. havilandii* Brandis (Ashton 1982, Saridan et al. 2013, Fambayun & Hut 2014). However, Ashton & Heckenhauer (2022) did not provide concise generic delimitation for the erected and/or new genera related to *Shorea* owing to many overlapping morphological characters (such as fruit calyx, stamen characters, and bark morphology) among the species. It is interesting to note that the Engkabang species, although now belonging to two genera, share similar fatty acids (Anderson 1975).

In recent years, several studies have been conducted on the genome of *Rubroshorea* (Heckenhauer et al. 2017, 2018 and 2019, Tian et al. 2022, Chew et al. 2023). The whole genome sequence of *R. leprosula* (published as

Shorea leprosula) and *R. macrophylla* have been published respectively by Ng et al. (2021) and Chung et al. (2025). In addition, several linkage maps have been constructed using different DNA markers such as randomly amplified polymorphic DNA (Lee et al. 2001), restriction fragment length polymorphisms (Indrioko et al. 2006), single-nucleotide polymorphisms (SNPs) (Heckenhauer et al. 2018), and simple sequence repeats (SSRs) (Ho et al. 2006, Chung et al. 2025). A genetic map with 101,066 SNPs has been reported by Heckenhauer et al. (2018). These DNA polymorphisms and genome sequences serve as effective tools for the identification of loci and genes responsible for nut traits.

In the past, the use of restriction-enzyme (RE)-based next-generation sequencing methods was established to study marker-assisted genetic studies. However, relatively large amounts of high-quality genomic DNA are required for the digestion steps. Enough amounts of DNA would be unavailable from small sized organisms such as microorganisms and field samples from, for example, endangered species. In ecological and conservation genetics, one would have to collect samples from degraded tissue which means RE-based NGS methods cannot be applied. Thus, a PCR-based procedure is introduced to tackle the limitation of RE-based NGS methods. Using multiplexed inter-simple sequence repeat (ISSR) primers, thousands of genome-wide regions were amplified effectively from a wide variety of genomes in the absence of a reference genome. This approach is called multiplexed ISSR genotyping by sequencing (MIG-seq) which is a PCR-based procedure for constructing highly reduced representation libraries without RE digestion steps. The approach allows for the discovery of de novo SNP and determination of the genetic distance between individuals which allows the construction of phylogenetic trees. The method is applicable to a wide range of marker-assisted genetic studies, particularly for medium-scale studies with less than 1,000 markers in ecological and conservation genetics. Thus, it is a quick survey of genetic differentiation among individuals of clone and breeding varieties to ascertain population of closely related species and hybrids (Suyama & Matsuki 2015).

In this study, we surveyed SNPs using the MIG-seq method from 96 accessions of *Rubroshorea* (*R. macrophylla* and *R. splendida*) and *Shorea* (*S. seminis*) collected from Central Sarawak and West Sarawak, Malaysian Borneo. The objectives of this study were to identify de novo genetic polymorphisms within the genome of the included accessions, to assess genetic similarity among the populations of *R. macrophylla*, and to compare the genetic differentiation among *R. macrophylla*, *R. splendida*, and *S. seminis*.

MATERIALS AND METHODS

Sample collection

A total of 96 accessions, comprising of 69 accessions of *R. macrophylla*, 14 accessions of *R. splendida*, and 13 accessions of *S. seminis*, were surveyed in this study (Table 1). These accessions were sourced from wild and semi-managed trees of various localities in West Sarawak and Central Sarawak, Malaysian Borneo. For the collection and preparation of leaf samples, the BioArk Leaf Collection Kit (Brand: LGC) was used to ensure secure and organised sample storage. To facilitate precise and contamination-free sample collection, a 6 mm Uni-Core Leaf Puncher (Brand: LGC) was employed in conjunction with a Leaf Punching Mat (Brand: LGC), providing a stable surface for consistent and efficient punching of leaf discs. This combination of tools ensured high-quality, standardised sample processing suitable for downstream genetic analyses. The prepared storage rack was placed inside the plastic bag. The package was shipped to LGC Berlin for the next step of the workflow.

Template DNA

Total DNA was extracted using suitable methods for various amounts of material ranging from 6 µg to 50 mg of each sample. Genomic DNA was extracted using the DNeasy Plant Mini Kit (QIAGEN, the Netherlands). The concentration and purity of DNA sample was assessed using Qubit 2.0 (ThermoFisher Scientific, Carlsbad, CA, USA) and NanoPhotometer N60 (Implen, Munich, Germany). Genome-wide SNPs were obtained by MIG-seq (Suyama & Matsuki 2015).

Primer selection

The MIG-seq primers were selected as following steps. For the two-base anchor sequences, only AC, AG, CC, GG, TC and TG combinations were used because C or G is suitable for 3′ ends of

Table 1 List of 96 *Rubroshorea* and *Shorea* accessions studied including species, collection site, and average size and weight of fruits

No.	Accession no.	Species	Locality	Nut size (length and width, cm)	Nut weight (g)
1	P01-A01-DIP59a	*R. macrophylla*	Sarawak, Serian, Pichin	6.0 x 4.0	51.0
2	P01-B01-DIP59b	*R. macrophylla*	Sarawak, Serian, Pichin	6.0 x 4.0	51.0
3	P01-C01-DIP59c	*R. macrophylla*	Sarawak, Serian, Pichin	6.0 x 4.0	51.0
4	P01-D01-DIP59d	*R. macrophylla*	Sarawak, Serian, Pichin	6.0 x 4.0	51.0
5	P01-E01-DIP59e	*R. macrophylla*	Sarawak, Serian, Pichin	6.0 x 4.0	51.0
6	P01-F01-DIP60a	*R. macrophylla*	Sarawak, Kuching, Grogo	–	–
7	P01-G01-DIP60b	*R. macrophylla*	Sarawak, Kuching, Grogo	–	–
8	P01-H01-DIP60c	*R. macrophylla*	Sarawak, Kuching, Grogo	–	–
9	P01-A02-DIP60d	*R. macrophylla*	Sarawak, Kuching, Grogo	–	–
10	P01-B02-DIP60e	*R. macrophylla*	Sarawak, Kuching, Grogo	–	–
11	P01-C02-DIP60f	*R. macrophylla*	Sarawak, Kuching, Grogo	–	–
12	P01-D02-DIP60g	*R. macrophylla*	Sarawak, Kuching, Grogo	–	–
13	P01-E02-DIP60h	*R. macrophylla*	Sarawak, Kuching, Grogo	–	–
14	P01-F02-DIP60i	*R. macrophylla*	Sarawak, Kuching, Grogo	–	–
15	P01-G02-DIP61a	*R. macrophylla*	Sarawak, Sri Aman, Sabal	–	–
16	P01-H02-DIP61b	*R. macrophylla*	Sarawak, Sri Aman, Sabal	–	–
17	P01-A03-DIP61c	*R. macrophylla*	Sarawak, Sri Aman, Sabal	–	–
18	P01-B03-DIP61d	*R. macrophylla*	Sarawak, Sri Aman, Sabal	–	–
19	P01-C03-DIP61e	*R. macrophylla*	Sarawak, Sri Aman, Sabal	–	–
20	P01-D03-DIP62a	*R. macrophylla*	Sarawak, Serian, Pichin	–	–
21	P01-E03-DIP62b	*R. macrophylla*	Sarawak, Serian, Pichin	–	–
22	P01-F03-DIP62c	*R. macrophylla*	Sarawak, Serian, Pichin	–	–
23	P01-G03-DIP62d	*R. macrophylla*	Sarawak, Serian, Pichin	–	–
24	P01-H03-DIP81a	*R. splendida*	Sarawak, Kuching, Grogo	5.5 x 3.0	23.0
25	P01-A04-DIP81b	*R. splendida*	Sarawak, Kuching, Grogo	5.5 x 3.0	23.0
26	P01-B04-DIP81c	*R. splendida*	Sarawak, Kuching, Grogo	5.5 x 3.0	23.0
27	P01-C04-DIP81d	*R. splendida*	Sarawak, Kuching, Grogo	5.5 x 3.0	23.0
28	P01-D04-DIP81e	*R. splendida*	Sarawak, Kuching, Grogo	5.5 x 3.0	23.0
29	P01-E04-DIP82a	*R. splendida*	Sarawak, Kuching, Opar	5.0 x 3.0	20.0
30	P01-F04-DIP82b	*R. splendida*	Sarawak, Kuching, Opar	5.0 x 3.0	20.0
31	P01-G04-DIP82c	*R. splendida*	Sarawak, Kuching, Opar	5.0 x 3.0	20.0
32	P01-H04-DIP82d	*R. splendida*	Sarawak, Kuching, Opar	5.0 x 3.0	20.0
33	P01-A05-DIP93a	*S. seminis*	Sarawak, Sri Aman, Sungai Raya	2.0 x 1.8	15.0
34	P01-B05-DIP93b	*S. seminis*	Sarawak, Sri Aman, Sungai Raya	2.0 x 1.8	15.0
35	P01-C05-DIP93c	*S. seminis*	Sarawak, Sri Aman, Sungai Raya	2.0 x 1.8	15.0
36	P01-D05-DIP93d	*S. seminis*	Sarawak, Sri Aman, Sungai Raya	2.0 x 1.8	15.0
37	P01-E05-DIP93e	*S. seminis*	Sarawak, Sri Aman, Sungai Raya	2.0 x 1.8	15.0
38	P01-F05-DIP105a	*R. macrophylla*	Sarawak, Kuching, Sungai Adis	6.0 x 5.0	55.0
39	P01-G05-DIP105b	*R. macrophylla*	Sarawak, Kuching, Sungai Adis	6.0 x 5.0	55.0
40	P01-H05-DIP105c	*R. macrophylla*	Sarawak, Kuching, Sungai Adis	6.0 x 5.0	55.0
41	P01-A06-DIP105d	*R. macrophylla*	Sarawak, Kuching, Sungai Adis	6.0 x 5.0	55.0
42	P01-B06-DIP105e	*R. macrophylla*	Sarawak, Kuching, Sungai Adis	6.0 x 5.0	55.0

No.	Accession no.	Species	Locality	Nut size (length and width, cm)	Nut weight (g)
43	P01-C06-DIP189a	*R. macrophylla*	Sarawak, Serian, Mongkos	–	–
44	P01-D06-DIP189b	*R. macrophylla*	Sarawak, Serian, Mongkos	–	–
45	P01-E06-DIP189c	*R. macrophylla*	Sarawak, Serian, Mongkos	–	–
46	P01-F06-DIP189d	*R. macrophylla*	Sarawak, Serian, Mongkos	–	–
47	P01-G06-DIP189e	*R. macrophylla*	Sarawak, Serian, Mongkos	–	–
48	P01-H06-DIP210a	*R. macrophylla*	Sarawak, Kuching, Bung Muan	–	–
49	P01-A07-DIP210b	*R. macrophylla*	Sarawak, Kuching, Bung Muan	–	–
50	P01-B07-DIP210c	*R. macrophylla*	Sarawak, Kuching, Bung Muan	–	–
51	P01-C07-DIP210d	*R. macrophylla*	Sarawak, Kuching, Bung Muan	–	–
52	P01-D07-DIP210e	*R. macrophylla*	Sarawak, Kuching, Bung Muan	–	–
53	P01-E07-DIO212a	*S. seminis*	Sarawak, Kuching, Grogo	–	–
54	P01-F07-DIP212b	*S. seminis*	Sarawak, Kuching, Grogo	–	–
55	P01-G07-DIP212c	*S. seminis*	Sarawak, Kuching, Grogo	–	–
56	P01-H07-DIP212d	*S. seminis*	Sarawak, Kuching, Grogo	–	–
57	P01-A08-DIP212e	*S. seminis*	Sarawak, Kuching, Grogo	–	–
58	P01-B08-DIP247a	*R. macrophylla*	Sarawak, Kuching, Krokong	–	–
59	P01-C08-DIP247b	*R. macrophylla*	Sarawak, Kuching, Krokong	–	–
60	P01-D08-DIP247c	*R. macrophylla*	Sarawak, Kuching, Krokong	–	–
61	P01-E08-DIP247d	*R. macrophylla*	Sarawak, Kuching, Krokong	–	–
62	P01-F08-DIP247e	*R. macrophylla*	Sarawak, Kuching, Krokong	–	–
63	P01-G08-DIP247f	*R. macrophylla*	Sarawak, Kuching, Krokong	–	–
64	P01-E09-DIP277a	*R. macrophylla*	Sarawak, Kapit, Sebabai	6.0 x 4.5	55.0
65	P01-F09-DIP277b	*R. macrophylla*	Sarawak, Kapit, Sebabai	6.0 x 4.5	55.0
66	P01-G09-DIP277c	*R. macrophylla*	Sarawak, Kapit, Sebabai	6.0 x 4.5	55.0
67	P01-H09-DIP277d	*R. macrophylla*	Sarawak, Kapit, Sebabai	6.0 x 4.5	55.0
68	P01-A10-DIP277e	*R. macrophylla*	Sarawak, Kapit, Sebabai	6.0 x 4.5	55.0
69	P01-B10-DIP277f	*R. macrophylla*	Sarawak, Kapit, Sebabai	6.0 x 4.5	55.0
70	P01-C10-DIP277g	*R. macrophylla*	Sarawak, Kapit, Sebabai	6.0 x 4.5	55.0
71	P01-D10-DIP277h	*R. macrophylla*	Sarawak, Kapit, Sebabai	6.0 x 4.5	55.0
72	P01-E10-DIP277i	*R. macrophylla*	Sarawak, Kapit, Sebabai	6.0 x 4.5	55.0
73	P01-F10-DIP277j	*R. macrophylla*	Sarawak, Kapit, Sebabai	6.0 x 4.5	55.0
74	P01-G10-DIP277k	*R. macrophylla*	Sarawak, Kapit, Sebabai	6.0 x 4.5	55.0
75	P01-H10-DIP277l	*R. macrophylla*	Sarawak, Kapit, Sebabai	6.0 x 4.5	55.0
76	P01-A11-DIP277m	*R. macrophylla*	Sarawak, Kapit, Sebabai	6.0 x 4.5	55.0
77	P01-B11-DIP277o	*R. macrophylla*	Sarawak, Kapit, Sebabai	6.0 x 4.5	55.0
78	P01-C11-DIP277p	*R. macrophylla*	Sarawak, Kapit, Sebabai	6.0 x 4.5	55.0
79	P01-D11-DIP278	*R. splendida*	Sarawak, Kapit, Sebabai	–	–
80	P01-E11-DIP279	*S. seminis*	Sarawak, Kapit, Sebabai	–	–
81	P01-F11-DIP280a	*R. macrophylla*	Sarawak, Sarikei, Julau	–	–
82	P01-G11-DIP280b	*R. macrophylla*	Sarawak, Sarikei, Julau	–	–
83	P01-H11-DIP280c	*R. macrophylla*	Sarawak, Sarikei, Julau	–	–
84	P01-A12-DIP280d	*R. macrophylla*	Sarawak, Sarikei, Julau	–	–
85	P01-B12-DIP281a	*S. seminis*	Sarawak, Sarikei, Julau	–	–
86	P01-C12-DIP281b	*S. seminis*	Sarawak, Sarikei, Julau	–	–

No.	Accession no.	Species	Locality	Nut size (length and width, cm)	Nut weight (g)
87	P01-D12-DIP282a	*R. macrophylla*	Sarawak, Sarikei, Pakan	–	–
88	P01-E12-DIP282b	*R. macrophylla*	Sarawak, Sarikei, Pakan	–	–
89	P01-F12-DIP282c	*R. macrophylla*	Sarawak, Sarikei, Pakan	–	–
90	P01-G12-DIP282d	*R. macrophylla*	Sarawak, Sarikei, Pakan	–	–
91	P01-H12-DIP282e	*R. macrophylla*	Sarawak, Sarikei, Pakan	–	–
92	P01-H08-DIP284a	*R. macrophylla*	Sarawak, Sarikei, Pakan	–	–
93	P01-A09-DIP284b	*R. macrophylla*	Sarawak, Sarikei, Pakan	–	–
94	P01-B09-DIP284c	*R. macrophylla*	Sarawak, Sarikei, Pakan	–	–
95	P01-C09-DIP284d	*R. macrophylla*	Sarawak, Sarikei, Pakan	–	–
96	P01-D09-DIP284e	*R. macrophylla*	Sarawak, Sarikei, Pakan	–	–

the primers. However, GC/ CG is not suitable because reverse complementarity. For the three-base core motifs, all combinations (4 x 3 x 2) were considered as candidates. Only half of them (reverse complement sequences) were treated as candidates for alternative set. For the two-base core motifs comprising two different bases, only AC/CA and their reverse complements TG/ GT, as an alternative set, were considered as candidates because AG/GA (polypurine repeat), TC/CT (polypyrimidine repeat), GC/CG (GC rich) and AT/TA (GC poor) are unsuitable for general primers. In the same way, only AAC/ CCA and their reverse complements TTG/GGT were considered as candidates for the three-base core motifs comprising two different bases. After combining the selected anchors with selected core motifs (e.g. 5′ -$(ACT)_4$TG-3′), primers that had: i) only repeated sequence (e.g. 5′ -$(AC)_6$AC-3′); ii) only one base difference from a simple sequence repeat (e.g. 5′ -$(AC)_6$AG-3′); iii) GC rich on three bases of the 3′ end (e.g. 5′ -$(AC)_6$CC-3′); iv) less than three different bases (comprising only two different bases) (e.g. 5′ -$(CA)_6$CC-3′) and; v) a reverse-complement sequence on four bases of its 3′ end (e.g. 5′-$(ACC)_4$GG-3′), were rejected as unsuitable. After combining with the tail sequences (e.g. 5′ -CGCTCTTCCGATCTCTG$(ACT)_4$TG-3′), primers that had more than three bases of reverse complement of the 3′ end sequence in its own or reverse primer were rejected as unsuitable. After comparisons with 'surviving' candidates in each alternative set, suitable primers for multiplexed PCR that did not have a reverse-complement conflict between three bases of the 3′ end and another primer sequence were selected. There were several choices to avoid the conflict; therefore, it was possible to select other primers instead of our choices. As a result, 8 × 8 and 4 × 4 primers were selected for the MIG-seq primer set-1 (Table 2) and set-2 (refer to Supplementary Table 2 in Suyama & Matsuki 2015).

Library construction

The first PCR step was performed to amplify ISSR regions from genomic DNA with MIG-seq primer set-1 (Table 2). Alternatively, MIG-seq primer set-2 (refer to Supplementary Table 2 in Suyama & Matsuki 2015) was used to create a different library from the same sample set. The volume of the PCR reaction mixture was 7 µL, containing 1 µL of template DNA, 0.2 µM of each first PCR primers, 3.5 µL of 2 × Multiplex PCR Buffer (Multiplex PCR Assay Kit Ver.2, Takara Bio, Kusatsu, Japan), and 0.035 µL of Multiplex PCR Enzyme Mix (Multiplex PCR Assay Kit Ver.2, Takara Bio). PCR was performed under the following conditions: initial activation at 94 °C for 1 min; 25 cycles for normal-concentration DNA (> 10 ng/µL) or 27 cycles for low-concentration DNA samples (< 5 ng/µL) of denaturation at 94 °C for 30 s, annealing at 48 °C for 1 min and extension at 72 °C for 1 min; followed by a final incubation at 72 °C for 10 min, using a GeneAmp PCR System 9700 (Thermo Fisher Scientific). The PCR products were visualized using a Microchip Electrophoresis System (MultiNA; Shimadzu, Kyoto, Japan) with the DNA-2500 Reagent Kit (Shimadzu).

The first PCR step could amplify a variety of ISSR regions including some mismatched priming sites, depending on the conditions,

Table 2 Sequences of MIG-seq primer set-1 for the first PCR. Underlined and boldface nucleotides denote tail and anchor sequences, respectively. The difference between the forward and reverse primer sets lies only in their tail sequences

Name	Sequences (5′–3′)
Forward primers: (Tail + **anchor: CTG**) + SSR + **anchor**	
$(ACT)_4$TG-f	CGCTCTTCCGATCT**CTG**ACTACTACTACT**TG**
$(CTA)_4$TG-f	CGCTCTTCCGATCT**CTG**CTACTACTACTA**TG**
$(TTG)_4$AC-f	CGCTCTTCCGATCT**CTG**TTGTTGTTGTTG**AC**
$(GTT)_4$CC-f	CGCTCTTCCGATCT**CTG**GTTGTTGTTGTT**CC**
$(GTT)_4$TC-f	CGCTCTTCCGATCT**CTG**GTTGTTGTTGTT**TC**
$(GTG)_4$AC-f	CGCTCTTCCGATCT**CTG**GTGGTGGTGGTG**AC**
$(GT)_6$TC-f	CGCTCTTCCGATCT**CTG**GTGTGTGTGTGT**TC**
$(TG)_6$AC-f	CGCTCTTCCGATCT**CTG**TGTGTGTGTGTG**AC**
Reverse primers: (Tail + **anchor: GAC**) + SSR + **anchor**	
$(ACT)_4$TG-r	TGCTCTTCCGATCT**GAC**ACTACTACTACT**TG**
$(CTA)_4$TG-r	TGCTCTTCCGATCT**GAC**CTACTACTACTA**TG**
$(TTG)_4$AC-r	TGCTCTTCCGATCT**GAC**TTGTTGTTGTTG**AC**
$(GTT)_4$CC-r	TGCTCTTCCGATCT**GAC**GTTGTTGTTGTT**CC**
$(GTT)_4$TC-r	TGCTCTTCCGATCT**GAC**GTTGTTGTTGTT**TC**
$(GTG)_4$AC-r	TGCTCTTCCGATCT**GAC**GTGGTGGTGGTG**AC**
$(GT)_6$TC-r	TGCTCTTCCGATCT**GAC**GTGTGTGTGTGT**TC**
$(TG)_6$AC-r	TGCTCTTCCGATCT**GAC**TGTGTGTGTGTG**AC**

SSR; simple sequence repeat (Suyama & Matsuki 2015)

because we decided to apply a low annealing temperature (48 °C in our recommended system) for the first PCR after checking several different temperatures. This annealing temperature could be effective to amplify more regions. The second PCR was performed to add the complementary sequences for the oligonucleotides that coat the Illumina sequencing flow cell, annealing sites of DNA sequencing primers, and indices to the first PCR products. Sequences of the common forward and indexed reverse primers were:5′-AATGATA CGGCGACCACCGAGATCTACACTCTTTC CCTACACGACGCTCTTCCGATCT CTG-3′ and 5′-CAAGCAGAAGACGGCATACGAGATxxxxxx GTGACTGGAGTTCAGACGTGTGCTCTT CCGATCTGAC-3′ where "xxxxxx" denotes the six-base index. This PCR step was conducted independently to add individual indices to each sample using the common forward and indexed reverse primers. The six-base index was designed using the Barcode Generator by Luca Comai and Tyson Howell (http://comailab.genomecenter.ucdavis.edu/index.php/Barcode_generator). The first PCR product from each sample was diluted 50 times with deionised water and used as the template for the second PCR.

The second PCR was performed in a 15-µL reaction mixture containing 3 µL of diluted 1st PCR product, 3 µL of 5 × PrimeSTAR GXL Buffer (Takara Bio), 200 µM of each dNTP, 0.375 U of PrimeSTAR GXL DNA Polymerase (Takara Bio), and 0.2 µM of common forward primer and individual reverse primer. The PCR conditions were as follows: 12 cycles of denaturation at 98 °C for 10 s, annealing at 54 °C for 15 s, and extension at 68 °C for 30 s. The concentrations of each second PCR product (libraries) were measured using a Microchip Electrophoresis System (MultiNA, Shimadzu) with a DNA-2500 Reagent Kit (Shimadzu). The libraries from each sample, each with a different index, were then pooled in equimolar concentrations. To reduce the salt concentration, the mixed libraries were purified, and the buffer was replaced with elution buffer using a QIAquick PCR Purification Kit (Qiagen, Venlo, Netherlands). Fragments in the size range of 300–800 bp in the purified library were isolated using Pippin

Prep DNA size selection system (Sage Science, Beverly, MA, USA). The final concentration was measured using a SYBR green quantitative PCR assay (Library Quantification Kit; Clontech Laboratories, Mountain View, CA, USA) with primers specific to the Illumina system.

Next generation sequencing

Libraries were denatured using fresh NaOH (0.2 N) and mixed with 10% of Illumina-generated PhiX control libraries, according to Illumina's protocol. Approximately 10 pM of the libraries were used for sequencing on NovaSeq 6000 Sequencer (Illumina), using a MiSeq Reagent Kit v3 (150 cycle, Illumina). Note that Illumina cluster generation algorithms are optimised to a balanced representation of A, C, G, and T nucleotides, and the sequence diversity of the first part of the sequencing is particularly critical for recognition of the cluster position in a flow cell. The first 17 bases of both ends of MIG-seq library (anchor and ISSR primer region) have biased nucleotides in the libraries; therefore, sequence reading of these nucleotides should be skipped to gain high-quality data. We skipped the sequencing of the first 17 bases of read 1 and three bases of read 2 (anchor region) using the 'DarkCycle' option of MIG-Seq system. We left the remaining 15 nucleotides in the reverse read (read 2), because the ISSR primer region can be used to gather information on which ISSR primers were used and does not cause problems in read 2 because it has no phasing step.

The procedure for setting the 'DarkCycle' is shown in Supplementary Figure 2 in Suyama & Matsuki (2015). Moreover, to replace the default index list with the original indices, a new 'SamplePrepkit' file was created, and the file name was added to the list of 'Compatible Sample Prep Kits' in the file of 'Assembly.txt' in 'Applications' folder in the Illumina Experiment Manager in MiSeq system (Supplementary Figure 3 in Suyama & Matsuki 2015). Both ends of the fragments and index sequences were read by paired-end sequencing (reads 1 and 2) and index sequencing; 80, 94 (except 'DarkCycle'), and six bases of sequences were determined as read 1, read 2, and the index read, respectively. In total, 180 bases were sequenced using a 150-cycle kit.

Data analysis

The data pre-processing phase involved several key steps to ensure high-quality sequencing data for downstream analysis. Initially, demultiplexing of all libraries for each sequencing lane was performed using Illumina bcl2fastq v2.20 software. During demultiplexing, one or two mismatches or ambiguous nucleotides (Ns) were allowed in the barcode reads when the barcode distances between all libraries on the lane made this permissible. Following this, sequencing adapter remnants were clipped from all reads, and those with a final length of less than 20 bases were discarded. Subsequently, MIG-seq primer sequences were clipped from the reads. In this step, up to three mismatches were allowed per primer, and both forward-reverse or reverse-forward primer pairs had to be present in each sequence fragment. After primer removal, sequences were reoriented into a forward-reverse orientation. Quality control was then assessed by generating FastQC reports for all FASTQ files. Finally, a summary spreadsheet titled read counts.xlsx was generated, providing a comprehensive overview of the read counts for all samples.

Clustering, alignment and SNP discovery

The process of clustering, alignment, and SNP discovery began with the clustering of overlap-combined reads using CD-HIT-EST v4.6.1, permitting up to 5% sequence difference. The resulting clusters were filtered to exclude singletons and clusters formed from fewer than 20 reads to ensure meaningful sequence groupings. Next, subsampled, quality-trimmed reads were aligned against the cluster reference using Bowtie2 v2.4.2, generating a single combined alignment file in coordinate-sorted BAM format for all samples. Variant discovery and genotyping were performed using Freebayes v1.2.0, with several stringent parameters applied, including minimum base quality of 10, minimum supporting allele quality sum of 10, a read mismatch limit of 3, minimum coverage of 5, and exclusion of indels, complex variants, and multi-nucleotide polymorphisms. Genotyping was conducted under diploid or tetraploid assumptions, and only observed genotypes were retained. Identified variants were further filtered using a MIG-seq-specific rule set, which required loci to have a minimum read count of 8, presence in at least 66% of samples, and a minimum allele frequency across all samples exceeding 5% or 10%. Finally, a genetic distance matrix was generated based on the Kosman/Leonard distance to facilitate analysis of genetic relationships among the samples.

Principal coordinate analysis

Pairwise distance matrix was calculated by counting the total number of alleles identical between any two accessions using MEGA ver. 7.0.21. Using the obtained distance matrix, principal coordinate analysis (PCoA) was conducted with the PCoA function in the stats package of PAST ver. 4.17 (Hammer et al. 2001).

Phylogenetic analysis

Phylogenetic analysis for all accessions was conducted using the Unweighted Pair Group Method with Arithmetic Average (UPGMA) cluster analysis in R ver. 4.5.0.

Morphological observations

Observation from field collections were made to record the nut size (length x width) and weight. Ten fruits were measured and weighed for each accession (where available).

RESULTS

Overview of sample quality

A total of 96 samples were sequenced. The number of reads per sample falls between 1.2 to 3.2 M reads. A minimum read count of 20 is recommended to accurately call genotypes in a diploid species with heterozygosity. Therefore, all samples passed the minimum read requirements. However, samples P01-B02-DIP60e, P01-G11-DIP280b and P01-H11-DIP280c have particularly small number of reads at 179000, 139000 and 213000, respectively. After further review, it was observed that many SNPs could not be determined for these three samples. Only 18%, 10% and 17% of SNPs, respectively, were observed for these samples. The percentage above is based on the number of SNPs reported over the total of number of SNPs observed. This is due to poor sample quality or DNA extraction.

De novo single nucleotide polymorphism discovery

The following discussion is based on two excel sheets (Supplementary Excel File 1: http://dx.doi.org/10.13140/RG.2.2.14127.85927 and Supplementary Excel File 2: http://dx.doi.org/10.13140/RG.2.2.25871.91049). Several filters were applied to determine significant polymorphisms in this study. The read count for a locus must exceed eight reads or abbreviated GMCF8 in the excel. The minimum number of samples (MinNS) refers to the minimum number of samples exhibiting any genotype. Minimum allele frequency (MinAF) is the frequency which the second most common allele occurs in a population. The minimum and maximum depth of coverage (MinMeanDP and MaxMeanDP) is 20 and 750 reads, respectively. This refers to the number of times a nucleotide is read during sequencing. The Supplementary Excel File 1 (http://dx.doi.org/10.13140/RG.2.2.14127.85927) is filtered with MinNS of 64 samples and MinAF of 0.1 while the Supplementary Excel File 2 (http://dx.doi.org/10.13140/RG.2.2.25871.91049) uses a MinNS of 48 samples with MinAF of 0.05. In Supplementary Excel File 1 (http://dx.doi.org/10.13140/RG.2.2.14127.85927), the number of SNPs reported is 421 polymorphisms with a range of allele frequency from 10% to 50%. The outcome of the filters in the Supplementary Excel File 2 (http://dx.doi.org/10.13140/RG.2.2.25871.91049) gave 1,160 polymorphisms. The observed allele frequency starts from 5% up to 50%.

Genetic distance matrix generation

Genetic distance measures the genetic divergence between species or between populations within a species, whether the measure of time from common ancestor or degree of differentiation. Populations with many similar alleles have small genetic distances, which indicates they are closely related and have a common ancestor. This study uses Kosman and Leonard distance as it is most suitable for populations with an asexual or mixed mode of reproduction, as the genetic frequencies and similarities between genotypes in the populations are considered. The genetic distance of this study is tabulated in the excel titled Supplementary Excel File 3 (http://dx.doi.org/10.13140/RG.2.2.35938.24002) and Supplementary Excel File 4 (http://dx.doi.org/10.13140/RG.2.2.19986.77766). The distance matrix among these accessions was exceptionally low, which may suggest close genetic similarity.

Principal coordinate analysis

Following the new taxonomy for the genus, *S. seminis* was separated from both species of *Rubroshorea*. Principal coordinate analysis diagram showed that populations of *R. macrophylla* are not separated according to localities (Figure 1).

Phylogenetic analysis

Phylogenetic trees were constructed to examine the genotype similarities and differences between populations (Figures 2 and 3). Phylogenetic analyses of 96 accessions were conducted using the common SNP sites. MIG-seq was able to

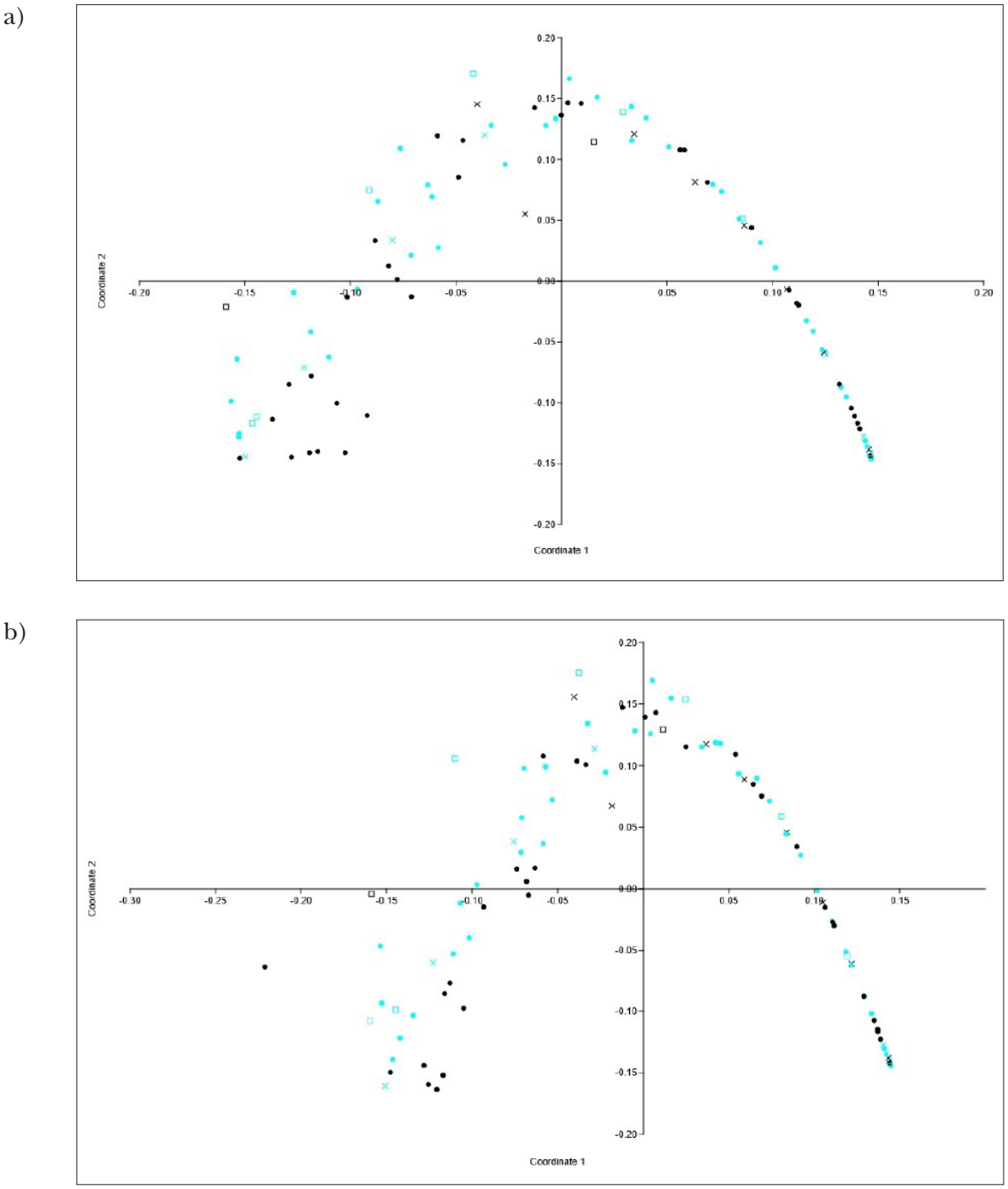

Figure 1 PCoA of *Rubroshorea* and *Shorea* accessions using distance matrix of (a) MinNS of 64 samples with MinAF of 0.1; (b) MinNS of 48 samples with MinAF of 0.05. The blue symbol denotes accessions from West Sarawak, while the black symbol indicates accessions from Central Sarawak. Dots represent *R. macrophylla* accessions, squares correspond to *R. splendida*, and crosses indicate *S. seminis* accessions

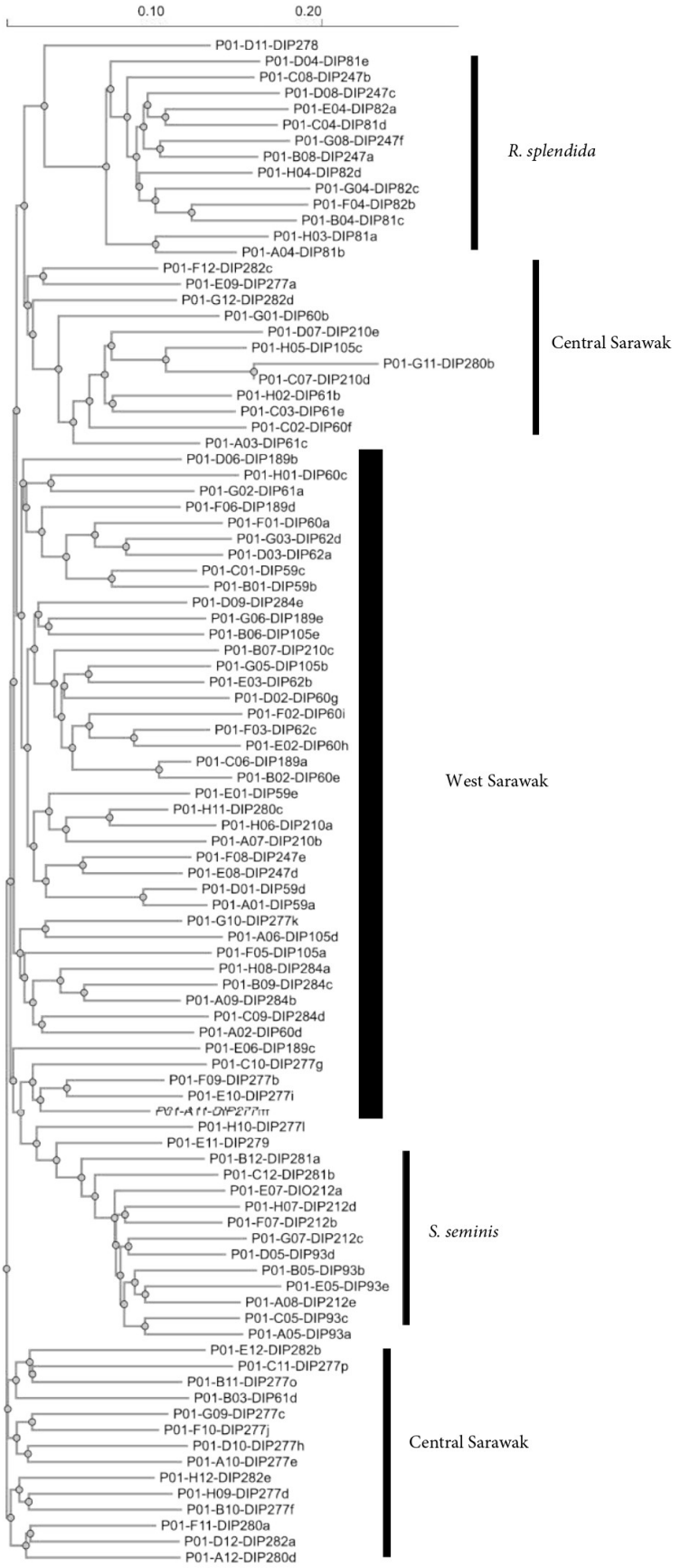

Figure 2 Phylogenetic tree of *Rubroshorea* and *Shorea* accessions using distance matrix of MinNS of 64 samples and MinAF of 0.1

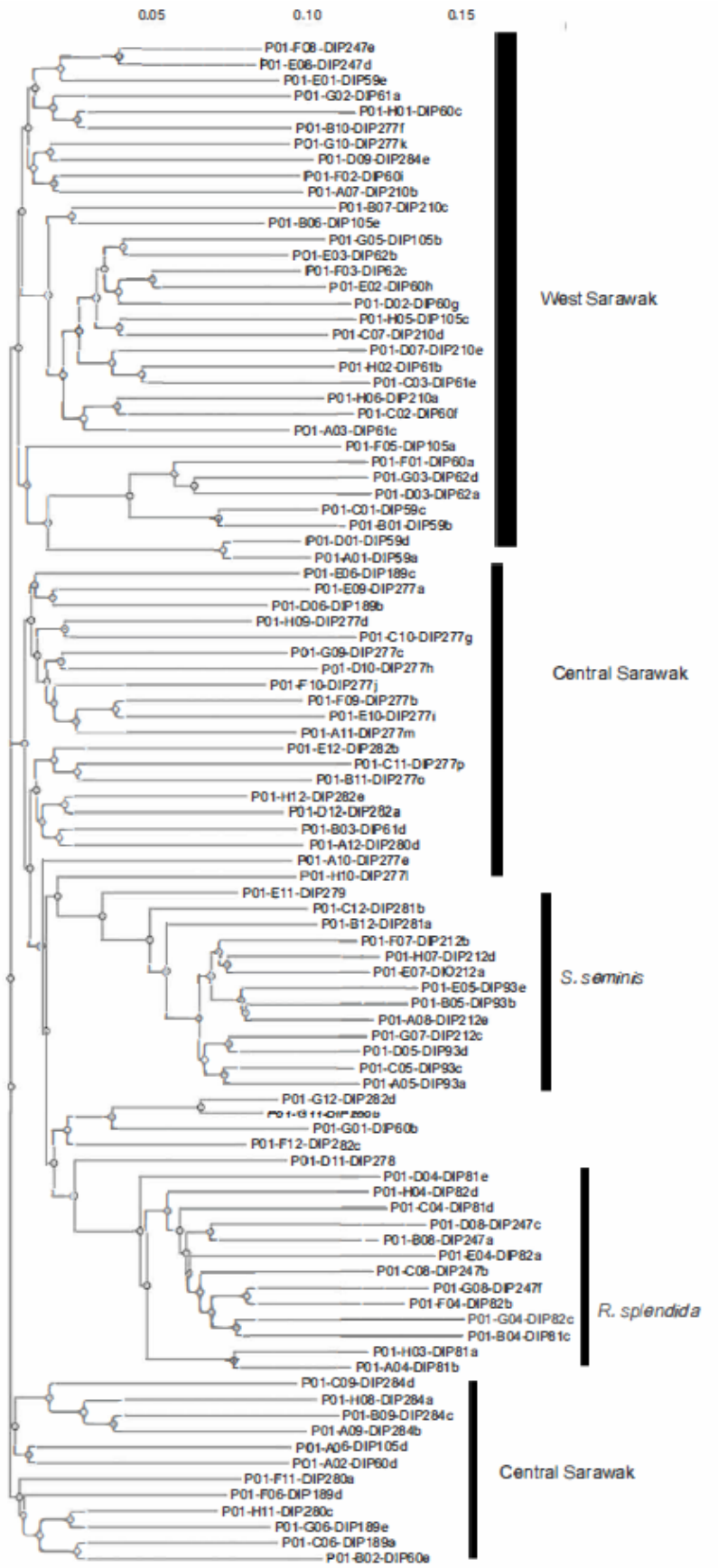

Figure 3 Phylogenetic tree of *Rubroshorea* and *Shorea* accessions using distance matrix of MinNS of 48 samples with MinAF of 0.05

discover new SNPs and construct a phylogenetic tree without the need of a reference genome. This is an extremely important advantage for new species with no prior reference genome and conservation studies where samples are often limited and scarce. The examined species and populations were divided into two major clades with *R. spendida* and *S. seminis* nested within the populations of *R. macrophylla* in Central Sarawak. Populations of *S. seminis* are nested together in a clade although the populations were sampled from various localities in Central and West Sarawak. Populations of *R. splendida* are placed together in the same clade.

Morphological observations

Fruits of *R. macrophylla* and *R. splendida* are much bigger than those of *S. seminis*, which concurs to previous studies (Ashton 1982, Yulita 2016). It is interesting to note that *S. seminis* has the most distinct shape of fruit wings and smaller fruit size (Figure 4). *Shorea seminis* is a riparian species at low altitude on mainly clay soil and is generally preferred for preparation for 'illipe butter' (Anderson 1975). However, the population of this species has greatly decreased owing to cutting of trees for timber usage (first author's personal observation), although the timber is of poor quality.

Figure 4 (a) *Rubroshorea macrophylla* fruit with radicle emerging (scale bar = 5 cm); (b) whole fruits of *R. macrophylla* for sale, Sibu (scale bar = 10 cm); (c) *Shorea seminis* fruit (scale bar = 3 cm); (d) peeled fruits of *S. seminis* for sale, Sarikei (scale bar = 2.5 cm)

DISCUSSION

This study revealed genome-wide DNA polymorphisms among *R. macrophylla*, *R. splendida*, and *S. seminis*. Although the exact number of loci was unknown, many genome-wide SNPs were identified among the 96 accessions. We employed more than 1000 informative, genome-wide SNPs filtered from MIG-seq from species in the tribe Shoreeae which are known to produce big nuts for oil production.

The phylogenetic trees obtained in the present study proved to be in agreement with their taxonomic identity at species level and localities (for *R. macrophylla*). Whereas *R. splendida* and *S. seminis* were nested within *R. macrophylla*, MIQ-seq data produce remarkably resolved phylogenomic trees. In addition, the relationships reported here represented the history along the genome, improving the confidence obtained from trees built from single loci like the plastid genome (Heckenhauer et al. 2019).

Knowledge of genome-wide DNA polymorphisms among these species could serve as a valuable resource for the genetics and breeding of *R. macrophylla* and its related species, and for understanding the evolution of the genus. Information on genome-wide SNPs among accessions belonging to these species is expected to significantly advance the genetics and breading of this species. Yulita (2016) is so far the only study which analysed 13 species which are known to produce Tengkawang nuts by using morphological and molecular methods. Their study showed that the tengkawang species indeed belonged to section *Pachycarpae* in *Shorea* (now in *Rubroshorea*) and the remaining species in *Shorea* as now defined.

The current study has provided compelling evidence that the included Engkabang species, which are mostly semi-managed by local villagers, are genetically closely related. By comparing thousands of SNP loci across individual trees, the study observed minimal genetic differentiation among these populations, showing an elevated level of genetic similarity. This close genetic relationship suggests that the semi-managed practices—such as selective harvesting and limited introduction of external genetic material—have supported a relatively homogeneous gene pool within these populations. The SNP marker data thus confirmed that despite potential environmental and management differences, the genetic makeup of these species remains highly conserved, underscoring the importance of these local practices in preserving their genetic integrity. Previous work on *R. macrophylla* (although the trees were planted) also confirmed that the species possessed moderate level of outcrossing rate and a significant level of inbreeding (Ng et al. 2002).

CONCLUSIONS

The discovery of informative SNPs in *R. macrophylla*, *R. splendida*, and *S. seminis* allows researchers to cross reference with previous studies of the Dipterocarpaceae family. By doing so, identification of SNPs unique to a species and SNPs shared across Dipterocarpaceae family can be identified. The genotypes found can be mapped to traits or phenotypes. Genotypes similarities and differences within and between populations can be confirmed by increasing the sample size from each population. These findings reinforced the value of SNP-based molecular tools in understanding the genetic structure of forest species and guiding sustainable management strategies.

ACKNOWLEDGEMENTS

Funding by the Sarawak Research and Development Council through Special Grant Scheme No. RDCRG02/THM02/2020/_1 is acknowledged. Fieldwork in Sarawak was conducted under Research Permit No. SFC.810-4/6/1(2023)-145 from Sarawak Forestry Corporation and SBC-2021-RDP-38-WSY from Sarawak Biodiversity Centre. This paper forms part of the results obtained from a PhD study of the second author.

SUPPLEMENTARY DATA

Supplementary data to this article can be obtained from the corresponding author:

Supplementary Excel File 1 Variants_GMCF8-HCT0.2_MinNS64_MinAF0.1_MinMeanDP20.0_MaxMeanDP750.0
http://dx.doi.org/10.13140/RG.2.2.14127.85927

Supplementary Excel File 2 Variants_GMCF8-HCT0.2_MinNS48_MinAF0.05
http://dx.doi.org/10.13140/RG.2.2.25871.91049

Supplementary Excel File 3 Distance_matrix_GMCF8-HCT0.2_MinNS64_MinAF0.1_MinMeanDP20.0_MaxMeanDP750.0
http://dx.doi.org/10.13140/RG.2.2.35938.24002

Supplementary Excel File 4 Distance_matrix_GMCF8-HCT0.2_MinNS48_MinAF0.05
http://dx.doi.org/10.13140/RG.2.2.19986.77766

REFERENCES

Anderson Jar. 1975. *Illipe Nuts (Shorea sp.) as Potential Agricultural Crops*. South East Asian Plant Genetic Resources, Bogor.

Ashton PS. 1982. Dipterocarpaceae. Pp 237–552 in Steenis (ed) *Flora Malesiana Series I, Vol. 9. Part 2*. National Herbarium of the Netherlands, Netherlands.

Ashton PS & Heckenhauer J. 2022. Tribe Shoreae (Dipterocarpaceae subfamily Dipterocarpoideae) finally dissected. *Kew Bulletin* 77: 885–903.

Blicher-Mathiesen U. 1994. Borneo Illipe, a fat product from different *Shorea* spp. (Dipterocarpaceae). *Economic Botany* 48: 231–242.

Chew Iyy, Chung HH, Lim Lwk et al. 2023. Complete chloroplast genome data of *Shorea macrophylla* (Engkabang): structural features, comparative, and phylogenetic analysis. *Data in Brief* 47: 109029.

Chung HH, Soh AAL, Lau MML, Gan HM, Sim SF & Lim LWK. 2025. The first engkabang jantong (*Rubroshorea macrophylla*) genome survey data. *Data in Brief* 58: 111248.

Coolen Q. 2013. *The Illipe Nut (Shorea spp.) as Additional Resource in Plantation Forestry. Case Study in Sarawak, Malaysia* (unpublished B.Sc. thesis). Van Hall Larenstein, University of Applied Sciences.

Darmawan MA, Muhammad BZ, Harahap AFP et al. 2020. Reduction of the acidity and peroxide numbers of tengkawang butter (*Shorea stenoptera*) using thermal and acid activated bentonites. *Heliyon* 6: e05742.

Fambayun RA & Hut S. 2014. *Budidaya Tengkawang Untuk Kayu Pertukangan, Bahan Makanan dan Kerajianan*. PT Penerbit IPB Press, Bogor.

Hammer Ø, Harper DAT & Ryan PD. 2001. PAST: Paleontological Statistics Software Package for Education and Data Analysis. *Palaeontologia Electronica* 4: 9 p. http://palaeo-electronica.org/2001_1/past/issue1_01.htm

Heckenhauer J, Paun O, Chase MW, Ashton PS, Kamariah AS & Samuel R. 2019. Molecular phylogenomics of the tribe Shoreeae (Dipterocarpaceae) using whole plastid genomes. *Annals of Botany* 123: 857–865.

Heckenhauer J, Samuel R, Ashton PS, Kamariah As & Paun O. 2018. Phylogenomics resolves evolutionary relationships and provides insights into floral evolution in the tribe Shoreeae (Dipterocarpaceae). *Molecular Phylogenetics and Evolution* 127: 1–13.

Heckenhauer J, Samuel R, Ashton PS et al. 2017. Phylogenetic analyses of plastid DNA suggest a different interpretation of morphological evolution than those used as the basis for previous classifications of Dipterocarpaceae (Malvales). *Botanical Journal of the Linnean Society* 185: 1–26.

Heri V, Bakara DO, Hermanto, Mulyana A, Moeliono M & Yuliani EL. 2020. Tengkawang sebagai 'perekat' pengelolaan daerah aliran Sungai terpadu. *Infobrief CIFOR* 292: 1–8.

Ho WS, Wickneswari R, Mahani MC & Shukor MN. 2006. Comparative genetic diversity studies of *Shorea curtisii* (Dipterocarpaceae): an assessment using SSR and DAMD markers. *Journal of Tropical Forest Science* 18: 22–35.

Kosman E. & Leonard KJ. 2007. Conceptual analysis of methods applied to assessment of diversity within and distance between populations with asexual or mixed mode of reproduction. *New Phytologist* 174: 683–696. https://nph.onlinelibrary.wiley.com/doi/10.1111/j.1469-8137.2007.02031.x

Lee SL, Wickneswari R, Mahani MC & Zakri AH. 2001. Comparative genetic diversity studies of *Shorea leprosula* (Dipterocarpaceae) using RAPD and allozyme markers. *Journal of Tropical Forest Science* 13: 202–215.

Ng KKS, Ang KC & Lee SL. 2002. Mating system in a planted population of *Shorea macrophylla* (Dipterocarpaceae). *Journal of Tropical Forest Science* 14: 145–149.

Ng KKS, Kobayashi MJ, Fawcett JA et al. 2021. The genome of *Shorea leprosula* (Dipterocarpaceae) highlights the ecological relevance of drought in aseasonal tropical rainforests. *Communications Biology* 4: 1166.

Saridan A, Fernandes A & Noor M. 2013. Sebaran dan potensi pohon tengkawang di hutan penelitian Labanan, Kalimantan Timur. *Jurnal Penelitian Dipterokarpa* 7: 101–108.

Smythies BE. 1958. The illipe nut. *The Sarawak Gazette* 84: 146–148.

Suyama Y. & Matsuki Y. 2015. MIG-seq: an effective PCR-based method for genome-wide single-nucleotide polymorphism genotyping using the next-generation sequencing platform. *Scientific Reports* 5: 16963. Link: https://www.nature.com/articles/srep16963

Tian Z, Zeng P, Lu X et al. 2022. Thirteen Dipterocarpoideae genomes provide insights into their evolution and borneol biosynthesis. *Plant Communications* 3: 100464.

Yulita KS. 2016. Phenetic and phylogenetic analyses of Tengkawang (*Shorea* spp., Dipterocarpaceae) based on morphological and molecular data. *Bulletin Kebun Raya* 19: 47–56.