Breeding Science 69: 1–10 (2019)
doi:10.1270/jsbbs.18061

Research Paper

SSR identification and marker development for sago palm based on NGS genome data

Devit Purwoko[1)], Imam Civi Cartealy[1)], Teuku Tajuddin[1)], Diny Dinarti[2)] and Sudarsono Sudarsono*[2)]

[1)] *Laboratory for Biotechnology, Agroindustrial Technology and Biotechnology, Agency for Assessment and Application of Technology*, Build. 630 Puspiptek Area Setu, South Tangerang 15314, Banten, Indonesia

[2)] *Plant Molecular Biology Laboratory, Department of Agronomy and Horticulture, Bogor Agricultural University*, Darmaga, Bogor 16680, West Java, Indonesia

Sago palm (*Metroxylon sagu* Rottb.) is one of the most productive carbohydrate-producing crops. Unfortunately, only limited information regarding sago palm genetics is available. This study aimed to develop simple sequence repeat (SSR) markers using sago palm NGS genomic data and use these markers to evaluate the genetic diversity of sago palm from Indonesia. *De novo* assembly of partial sago palm genomic data and subsequent SSR mining identified 29,953 contigs containing 31,659 perfect SSR loci and 31,578 contigs with 33,576 imperfect SSR loci. The perfect SSR loci density was 132.57/Mb, and AG, AAG and AAAT were the most frequent SSR motifs. Five hundred perfect SSR loci were randomly selected and used for designing SSR primers; 93 SSR primer pairs were identified. After synteny analysis using rice genome sequences, 20 primer pairs were validated using 11 sago palm accessions, and seven primers generated polymorphic alleles. Genetic diversity analysis of 41 sago palm accessions from across Indonesia using polymorphic SSR loci indicated the presence of three clusters. These results demonstrated the success of SSR identification and marker development for sago palm based on NGS genome data, which can be further used for assisting sago palm breeding in the future.

Key Words: *Metroxylon sagu*, genome sequencing, SSR mining, microsatellites, SSRs.

Introduction

Sago palm (*Metroxylon sagu* Rottb.) is one of the most productive carbohydrate-yielding plants worldwide (Ishizaki 1997). A sago palm tree can approximately accumulate 100–300 kg starch in its trunk (Dewi *et al*. 2016). Moreover, sago palm yields four times more starch than rice (Karim *et al*. 2008). Therefore, sago palm is a potential solution for the impending worldwide food crisis (Abbas *et al*. 2010). This monocot species has a diploid number of 26 chromosomes (2n = 2x = 26). It belongs to family *Arecaceae* and order *Arecales* (Flach 1997). Sago palm occurs throughout the Southeast Asian region. However, Beccari (1918) proposed the Maluku Islands in the eastern part of Indonesia as the centre of sago palm genetic diversity. Sago palm is relatively tolerant of abiotic stress environments, especially that of swampy and waterlogged areas (Singhal *et al*. 2008). Moreover, sago palm can thrive in acidic peat soils with high concentrations of metal compounds (Miyamoto *et al*. 2009). Most current commercial crops are unlikely to survive under such suboptimum conditions (Tajuddin *et al*. 2007).

Although sago palm is an essential starch-producing crop and has excellent environmental adaptability (Tajuddin *et al*. 2007, Uthumporn *et al*. 2014, Wee and Roslan 2012), until recently, attention to this crop has been insufficient (Karim *et al*. 2008). Most commercial producers only harvest sago palm from natural sago forests and invest little for sustainable use. Although phenotypic variabilities exist among natural populations of sago palms, breeding for specific phenotypes and traits is still necessary. Unfortunately, understanding of sago palm genetics is also limited (Wee and Roslan 2012). Therefore, modern molecular biology approaches are required to support further elucidation of sago palm biology and genetics in order to support a large-scale cultivatation of sago palm.

Some researchers have reported the use of molecular markers for elucidating genetic information in sago palm.

Communicated by Sachiko Isobe
Received May 11, 2018. Accepted September 26, 2018.
First Published Online in J-STAGE on March 16, 2019.
*Corresponding author (e-mail: sudarsono.ipb@gmail.com)

Abbreviations: bp: base pair, Gb: Gigabase-pair, Mb: Megabase-pair, NGS: Next-generation sequencing, PIC: Polymorphism information content, SSR: Simple sequence repeat

However, most of these reports have described a limited number of marker loci, for either sago provenance or population samples, or both. Abbas *et al*. (2009) used RAPD markers to study sago palm genetics. In other studies, Kjær *et al*. (2004) used AFLPs and Abbas *et al*. (2010) used chloroplast DNA (cpDNA) markers for studying the crop. The availability of more robust molecular markers will assist in improving the understanding of sago palm genetics.

Understanding the genetic variability of natural sago palm populations is necessary to support future sago palm breeding (Abbas *et al*. 2010) and genetic resource conservation programmes (Kjær *et al*. 2004); however, robust genetic markers must be developed to support these important endeavours. Unfortunately, markers capable of providing high-resolution information regarding sago palm genome have not been readily available.

Simple sequence repeats (SSRs), or microsatellites, are repetitive DNA sequences that consist of 1–6 bp motifs widespread across both prokaryotic and eukaryotic genomes (Grover *et al*. 2012, Guo *et al*. 2009, Kelkar *et al*. 2008, Sharma *et al*. 2007). Such SSRs are useful for molecular marker development because they are abundant, highly polymorphic, multiallelic and inherited codominantly (Singh Kesawat and Das 2009). SSRs have been widely employed as genetic markers for many genetic studies on crops (Ashkani *et al*. 2012, Geethanjali *et al*. 2017, Girichev *et al*. 2017, Kaur *et al*. 2016, Ott *et al*. 2011, Rauscher and Simko 2013, Zong *et al*. 2015). Despite the many advantages of SSR markers, one disadvantage is the high cost of marker development because it requires extensive sequencing of the target plant genome (Zalapa *et al*. 2012). This constraint hinders the usage of SSR markers for crops with limited genome sequence information, such as sago palm.

Recent developments in DNA sequencing technology, such as next-generation sequencing (NGS), have offered a new avenue to acquire large genome sequences of non-model crops, mine SSR marker sequences and develop the required primers to generate SSR markers (Zalapa *et al*. 2012). Many examples of the use of such genomic data for SSR marker development in non-model crops have recently become available in the literature (Zalapa *et al*. 2012). Although some palm species have received great attention using NGS sequencing technology, unfortunately, sago palm has not.

In this study, we aimed to identify and develop new SSR markers from partial genome sequences of sago palm using the Illumina GAIIx platform and paired-end genomic fragment libraries. Following *de novo* assembly of the raw reads, the partial genomic sequence data were subsequently used to mine SSR sequences, design appropriate primers and develop specific SSR markers for sago palm. After validation of the developed markers, some polymorphic markers were used to evaluate the genetic diversity of 41 Indonesian sago palm accessions. To the best of our knowledge, this paper is the first to report SSR marker mining using genomic sequences of sago palm and use generated markers for evaluating diverse Indonesian sago palm accessions.

Materials and Methods

Plant materials

For genome sequencing and SSR marker development, we utilised the collection of sago palm accessions of the Indonesian Agency for Assessment and Application of Technology (BPPT). For initial SSR marker validation, we used 11 sago palm accessions. Subsequently, we used a diverse assortment of 41 sago palm accessions originating from different regions in Indonesia for genetic diversity studies (**Fig. 1**). Fresh leaf samples were collected from BPPT sago palm accessions representing all available provenances and used for DNA isolation. Additionally, fresh leaf samples from field points of origin of sago palm were collected,

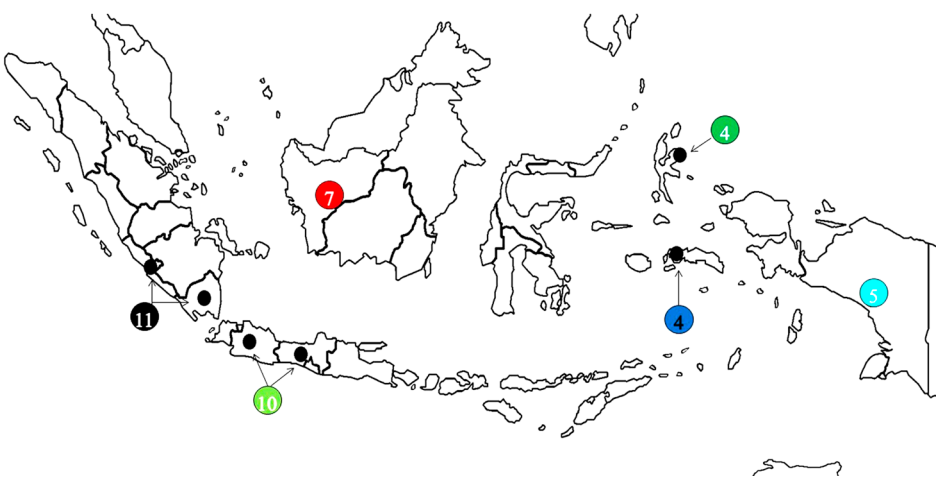

Fig. 1. Map illustrating the origin of sago palm samples used to analyse genetic diversity using SSR markers. Sago palm accessions originated from (●) Sumatra, (●) Seram, (●) Java, (●) Borneo, (●) Halmahera and (●) Papua. The number in the circles indicates the number of accessions collected from each location.

wrapped in paper and sent by airmail to the Biotechnology Lab, BPPT, for DNA isolation.

Total DNA isolation, library preparation, NGS and genome assembly

Total DNA was isolated from leaf samples of sago palm using the standard CTAB method modified for DNA isolation from palm leaves (Maskromo *et al*. 2016, Novero *et al*. 2012, Pesik *et al*. 2015, 2017, Tinche *et al*. 2014). We performed sago palm genome sequencing of DNA extracted from young leaves using the Illumina GAIIx instrument. Paired-end genomic library construction (2 × 72 bp) was conducted with a commercial Nextera XT Index with TruSeq Dual Index Sequencing Primer Box kit, following the manufacturer's protocol (https://www.illumina.com/products/by-type/sequencing-kits/cluster-gen-sequencing-reagents/truseq-dual-index-seq-primers.html). After quality trimming of raw reads with Trimmomatic, we used Ray software (https://github.com/sebhtml/ray) for *de novo* genome assembly. Subsequently, we used the assembled sago palm genome sequences for SSR mining and marker development.

SSR sequence mining from partial sago palm genome sequence

We used the assembled contig data (at least 200 bp) to search for di-, tri-, tetra- and hexanucleotide repeats of SSR loci of at least 20 bp lengths. We used Phobos software (http://www.ruhr-uni-bochum.de/ecoevo/cm/cm_phobos.htm) to mine SSR motifs from the assembled genome sequence. Identified SSR loci were then grouped into either perfect or imperfect SSRs and designated as either class I or II SSRs. SSR locus density was determined based on the frequency of SSR loci and the total length of contigs containing SSRs. We also evaluated the motif length, loci numbers, mean repeat numbers and densities for the selected repetitive motifs.

SSR primer design and primer validation

To design SSR primers, we selected 500 of the total class I SSR loci with a minimum of 10-fold coverage from the outputs of Phobos software. The contigs containing selected SSRs were used for SSR primer design using Primer3-Plus software (http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi). The parameters for SSR primer design included 200–600 bp amplicon size, 18 bp optimum primer size, 50°C–60°C primer melting temperature (Tm) and 40%–60% primer GC content.

Once the SSR primers were identified, we performed synteny analysis on selected contigs containing SSR loci. Synteny analysis was performed on rice (*Oryza sativa*) chromosome sequences available from the Phytozome website (https://phytozome.jgi.doe.gov/pz/portal.html#!search?show=BLAST) to evaluate their probable position distributions in the rice genome. Subsequently, we selected 20 primer pairs distributed across the 11 rice chromosomes for primer validation. The ability of the selected SSR primer pairs to amplify polymorphic markers was evaluated and validated using 11 sago palm accessions. The SSR primers successfully yielded polymorphic markers across the 11 sago palm accessions during primer validation steps and were subsequently used for genetic diversity analysis.

PCR amplification and allele identification

PCR amplification mixtures consisted of 5 µL of 5× Taq Polymerase buffer, 0.25 µL KAPA Taq HotStart Extra (100 units/µL), 1.5 µL $MgCl_2$ (25 mM), 0.5 µL dNTP (10 mM), 0.5 µL Forward and Reverse primers (100 mM) and made up to 25 µL with sterile ddH_2O. The Takara PCR Thermal Cycler Dice® (http://catalog.takara-bio.co.jp/product/basic_info.php?unitid=U100004192) was used for SSR marker amplification. First, DNA extract was subjected to one cycle of denaturation at 95°C for 3 min. This was followed by 35 cycles of denaturation at 95°C for 30 s, primer annealing at the appropriate Tm for each primer pair for 30 s and primer extension at 72°C for 30 s. Finally, there was a final extension step at 72°C for 60 s.

For SSR allele identification, we used denaturing polyacrylamide gel electrophoresis (PAGE) using a vertical slab gel DNA sequencer (34 × 45 cm) in a 6% SB (1×) buffer-polyacrylamide gel (Brody and Kern 2004). Allele visualisation employed the silver staining method, as described by Chevallet *et al*. (2016). We scored markers manually and selected the polymorphic ones for genetic analysis.

Sago palm genetic diversity and structure assessment

We calculated a dissimilarity matrix based on allelic data for the diploid using a simple matching dissimilarity index. The calculation of dissimilarity matrices used bootstrap analysis with 10,000 iterations. Principal coordinate analysis (PCoA) based on dissimilarity was set using the option of 41 axes to edit, and the default axis as determined by the PCoA was selected. We performed tree construction using the calculated dissimilarity matrix by the weighted neighbour-joining approach. The dissimilarity matrix, bootstrapping, PCoA and tree construction for the sago palm accessions were conducted using Dissimilarity Analysis and Representation for WINDOWS (DARWin) software version 6.05 (Perrier and Jacquemoud-Collet 2006; http://darwin.cirad.fr/darwin).

We calculated population genetic parameters (allele numbers, He, Ho and PIC) for each of the SSR marker loci using CERVUS software version 3.0 (Kalinowski *et al.* 2007) and GENALEX software version 6.501 (Peakall and Smouse 2012). STRUCTURE software version 2.3.4 (Pritchard *et al*. 2000, http://pritch.bsd.uchicago.edu/structure.html) was used to analyse population structures and differentiate allele frequencies. For calculations to estimate an ideal number of populations (K), we ran each of the K estimate in an admixture model, with K = 1–10 and with each K replicated 20 times. We implemented each replication with a burn-in period of 100,000 steps followed by 250,000 replications of

Monte Carlo Markov Chain model generation. *Ad-hoc* statistics were evaluated to estimate changes in the log probability of data according to the K value, as suggested by Evanno *et al.* (2005). The ideal number of population clusters was determined based on the highest K value as estimated using STRUCTURE HARVESTER (http://taylor0.biology.ucla.edu/struct_harvest/) (Earl and vonHoldt 2012).

Results

NGS and assembly of partial sago palm genome

In this study, a total of 315.56 MB of raw reads of partial sago genome sequence data was generated using the Illumina GAIIx paired-end NGS system (**Table 1**). Results of total nucleotide composition analysis indicated that adenine was the most frequent base (A = 31.4%), followed by thymine (T = 30.7%), cytosine (C = 18.4%) and guanine (G = 18.3%). The percentage of GC content in partial genome sequences of sago palm was approximately 37%. Following *de novo* assembly, we identified a total of 904,670 contigs (**Table 1**). The minimum length of the assembled contigs was 100 bp and the maximum was 355,487 bp. The average length of the assembled contig was 263 bp, whereas that of N50 was 291 bp (**Table 1**).

SSR sequence mining

We identified 29,953 contigs containing 31,659 loci of perfect SSRs and 31,578 contigs with 33,576 loci of imperfect SSRs (**Table 2**). Further analysis also indicated that 12,673 (40.03%) SSR loci were class I SSRs (repeat length ≥20 bp), with a density of 40.2 SSR/Mb, and 18,986 (59.97%) were class II SSRs (repeat length 12–20 bp), with a density of 60.2 SSR/Mb (**Supplemental Fig. 1A**).

Dinucleotides (17,376; 55%) were the most frequently found types among perfect SSR motifs, and hexanucleotides (2,196; 7%) were the least frequently found (**Table 3**). The cumulative length of the SSR-containing contigs was 335,162 kb for dinucleotide and 45,885 kb for hexanucleotide SSRs, while the SSR loci densities were 72.06 loci/Mb for dinucleotides and 9.19 loci/Mb for hexanucleotides (**Table 3**). The frequency, SSR cumulative length and density of SSR loci occurrence in the partial sago palm genome decreased with increasing motif length (di- to hexanucleotides, **Table 3**). The mean repeat number for dinucleotides was 9.6, whereas that for hexanucleotides was 3.5 (**Table 3**).

The most frequently found sequence motifs for the class I and class II SSR loci were AG, AAG or AAAT. Among the four dinucleotide repeat motifs found in the sago palm genome (AC, AG, AT or CG), the AG repeat was most common (38.7%), whereas the CG repeat was least common (0.22% of the total number of SSRs). For trinucleotides, the AAG repeat (5.2%) was most common, whereas the ACG repeat was least common (0.12% of the total SSR). We did not identify any CCG trinucleotide repeats in the results of the NGS SSR sequence mining of the sago palm genome (**Supplemental Fig. 1B**, **1C**).

SSR primer design and primer validation

To design PCR primers, we randomly selected 500 of 31,659 identified class I SSR loci. However, for 407 (81.4%) of these randomly selected loci, we could not design the flanking primers because of either unsuitable flanking sequences or Tm constraints. We designed flanking primers for 93 (18.6%) selected loci that consisted of 37 dinucleotide, 24 trinucleotide and 32 tetranucleotide repeats (**Supplemental Table 1**). The contig sequences used to design primers were deposited in the NCBI GenBank DNA Database under the accession no. MG904300-MG904384.

Results of synteny analysis using rice genome data indicated that 55 of the 93 selected SSR loci were represented on the 11 rice chromosomes; therefore, 38 sago palm loci were not found within any rice chromosome. For primer validation, we selected 16 SSR primer pairs distributed across the 11 rice chromosomes and four pairs of unknown location (**Table 4**). Selected primer pairs amplified nine loci of dinucleotide, six of trinucleotide and five of tetranucleotide repeats. The results of primer validation (**Supplemental Fig. 2A**) indicated that only seven primer pairs produced

Table 1. Next Generation Sequencing (NGS) and *de novo* assembly result summary from partial sago palm (*Metroxylon sagu*) genome sequences

NGS and *de novo* assembly summary	Number
Reads sequences (Mb)	315,563,284
Contig sequence numbers	904,670
Contigs length (bp)	238,803,664
Shortest contig (bp)	100
Longest contig (bp)	355,487
Mean length of contigs (bp)	263
N50	291

Table 2. SSR sequence mining result summary from partial sago palm (*Metroxylon sagu*) genome sequences

SSR sequence mining summary	Number
Total SSR number :	
Perfect	31,659
Imperfect	33,576
Contig containing SSR:	
Perfect	29,953
Imperfect	31,578

Table 3. Distribution of perfect SSRs in the genomic sequences of sago palm (*Metroxylon sagu*)

Motifs length	Number of loci identified	Mean of repeat number	Cumulative length (kb)	Density (SSR/Mb)*
Di-	17,376	9.6	335,162	72.76
Tri-	5,682	6.1	104,216	23.79
Tetra-	3,816	4.8	72,671	15.98
Penta-	2,589	4.0	51,211	10.84
Hexa-	2,196	3.5	45,885	9.19
Total	31,659	27.9	609,145	132.57

*Density of SSR was calculated using ratio between the number of SSR loci over the identified total contig length (238.80 Mb).

Table 4. List of sequences of SSR primer pairs used in the validation of SSR markers

No.	Primer	Primer sequences (5′-3′)	Motifs	Product size (bp)	Results of synteni analysis	
					Contig location in rice chromosome	E-value
1	sV16071	F: TGCCACTGGTGAAGAGCA R: TTCTCGAGGCCGTTCTTG	AAGG	497	Chr 3	7.00E-75
2	sV4223	F: TCATCAGCCCCTTCAGATG R: CACGCTGAGGCAGAGAAA	ATC	444	Not found	Not found
3	sV523089	F: TCCCAAAAGGGCAAACAA R: AGAAAAGTCTGGGCAGATCG	AAG	381	Not found	Not found
4	sV7173	F: TGCTGGTTCTCTTGTCGTGT R: TCTCCCCTCCGGACATTT	AG	419	Chr 11	0.002
5	sV4442	F: CATGCATGCACTGTTTGCT R: GAGCGTTGGTTGCTCGAT	AAT	433	Chr 1	1.00E-15
6	sV196074	F: TGACCGAGGCAAGCTAGTG R: AGCTTGCGTGTTGCATTG	AAAG	405	Not found	Not found
7	sV646	F: GTAGCTGATTGCCCACTTAC R: ATGGCACCACATCTTCTAAC	AGC	229	Chr 1	7.00E-19
8	sV95916	F: GGCATGCCCTATACAATTAC R: TGTGCCTTGCATGTATAAAG	ATCC	233	Not found	Not found
9	sV7446	F: CCTTCAGATAAACTGGTGGA R: CTCCTCGTAACAGAGAGGTG	AG	230	Chr 12	5.00E-09
10	sV2006	F: GTATAGATGGAAAGCGTTGG R: CCGCTCCTTATCCTAGTCTT	AT	247	Chr 2	3.00E-28
11	sV400785	F: ACTCCGCTCACTTGCACA R: GCACGCCTAAGGATGGAA	AG	300	Chr 5	8.00E-08
12	sV513907	F: GGCGGAGCTTCAAGAACA R: TCAATGCCAGACAAAGATGC	AG	312	Chr 6	0.016
13	sV67385	F: AGCACCGAAGGAAACAACC R: AGCCGAAAAGCCGAGTCT	AG	310	Chr 7	1.00E-05
14	sV6886	F: GACATGCTTGGCCTTGGT R: CCTTGGTTGGAACCCTCA	AT	448	Chr 8	9.00E-07
15	sV109470	F: CCCATGCCTTATGCTGGA R: CTTGCTGGCTAGTGCCAAT	AAG	360	Chr 9	4.00E-24
16	sV100242	F: TTGAGCCAGGTATCATCCAA R: ATCGTGGCAGAAGGTGGT	AAAC	308	Chr 10	0.028
17	sV2283	F: ACGGACCAGTCGGCATTA R: TCGGGGAGAGAGCGATTA	AG	596	Chr 2	2.00E-63
18	sV328094	F: AACTGATGGGTGGGCAAA R: GCATGCACATGGGAGACA	AG	471	Chr 3	3.00E-47
19	sV72_1	F: TCAGCCTTCCCTTCCTCA R: ACAGCACATCGCAAGCAC	AAG	564	Chr 5	9.00E-05
20	sV897681	F: AGCACCGCGTGGAAAGTT R: GCAACACATCTCCCACCA	AAAT	453	Chr 6	0.011

polymorphic SSR markers, 12 produced monomorphic markers and one primer pair failed to generate any amplicon across the 11 sago palm accessions investigated.

Sago palm genetic diversity and structure assessment

Results of the analysis using seven polymorphic SSR marker loci across 41 sago palm accessions indicated that there were 2–5 alleles per locus, with an average of 3.4 alleles (**Table 5**; see **Supplemental Fig. 2B** for a representative sample of silver stained acrylamide gel showing polymorphic SSR alleles). The estimated polymorphic information content (PIC) ranged from 0.356 to 0.704, with an average of 0.475 (**Table 5**). The sV2006 SSR locus (see **Table 5**) yielded the highest number of alleles (5) and highest PIC (0.704; **Table 5**).

Expected heterozygosity (He), estimated using each of the evaluated SSR markers among 41 sago palm accessions, ranged from 0.459 to 0.758, with an average of 0.579. However, the observed homozygosity (Ho), estimated using each of the SSR markers, ranged from 0.293 to 0.878, with an average of 0.631. The sV2006 SSR locus exhibited the highest He, whereas sV2283 exhibited the lowest. However,

Table 5. Summary of observed allele number (N), polymorphism information content (PIC), observed and expected heterozygosity (Ho and He) for 53 sago palm accession

No.	SSR Loci ID	Estimated allele size (bp)	N	PIC	Ho	He
1	sV2006	247–350	5	0.704	0.780	0.758
2	sV400785	300–620	3	0.530	0.854	0.604
3	sV513907	312	3	0.406	0.561	0.522
4	sV67385	310–410	4	0.533	0.878	0.620
5	sV109470	360–400	3	0.390	0.366	0.494
6	sV100242	308	2	0.356	0.683	0.470
7	sV2283	596	4	0.406	0.293	0.459
Average			3.429	0.475	0.631	0.579

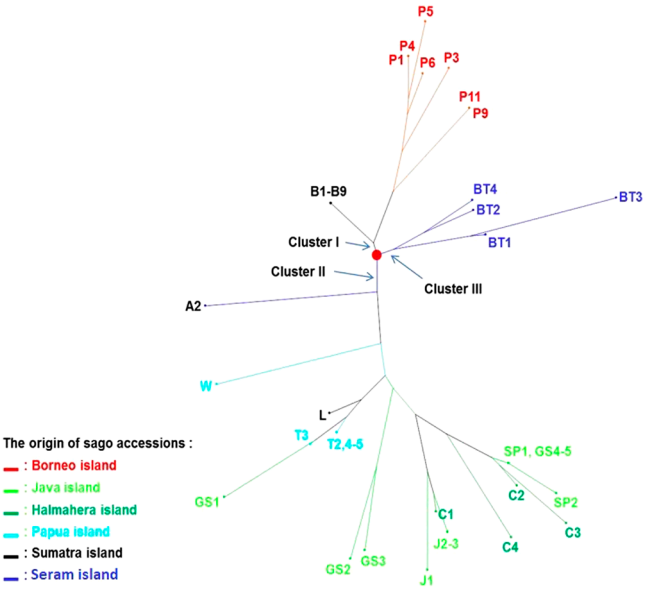

Fig. 2. Unrooted weighted neighbour-joining cluster analysis of genetic dissimilarity as measured using amplified simple sequence repeat (SSR) markers. Accessions and collection localities are indicated in colour labels.

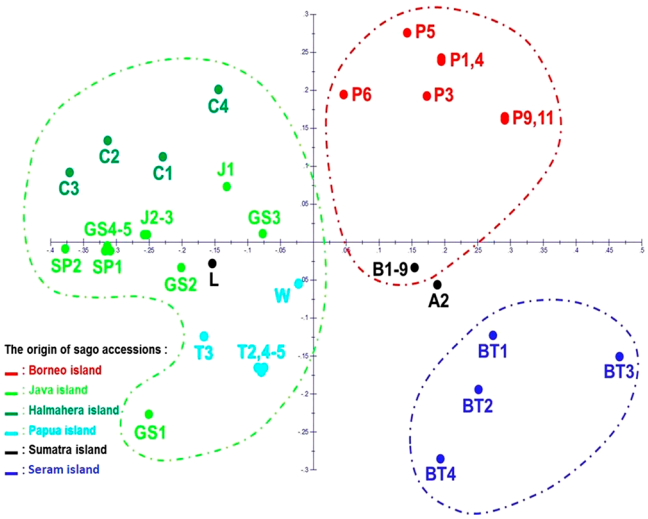

Fig. 3. Factorial analysis based on Eigen values calculated from seven SSR markers. The 41 sago palm accessions were clustered into three populations represented here by different colours.

the sV67385 SSR locus exhibited the highest Ho, whereas sV2283 exhibited the lowest.

Unrooted weighted neighbour-joining cluster analysis for the 41 sago palm accessions using DARWin software grouped the accessions into three clusters (**Fig. 2**). The first cluster (Cluster I) consisted of nine sago accessions from Sumatra and seven from Borneo; the second cluster (Cluster II) consisted of 10 sago palms from Java, four from Halmahera, five from Papua and one from Sumatra (L) and the third cluster (Cluster III) consisted of four sago palms from Seram and one from Sumatra (A2) (**Fig. 2**). Based on their phenotypes, BT1, BT2, BT3 and BT4 accessions (from Seram), A2 (from Sumatra), C4 (from Halmahera) and W (from Papua) were all of the spiny type, whereas the remaining accessions were spineless.

The results of PCoA (**Fig. 3**) presented a two-dimensional graphical view of the genetic diversity of 41 sago palm accessions originating from various regions in Indonesia. The clustering of sago accessions from PCoA supported dendrogram cluster analysis but not genetic structural analysis. The dendrogram cluster analysis grouped the sago accessions into three major groups (**Fig. 2**), and the genetic structural analysis grouped them into two major groups (**Fig. 4**).

STRUCTURE V2.3.4 was used for genetic structural analysis and population distribution. Simulations were run with 100,000 iterations and population number (K) ranging from 1 to 10. Each K value was run over 20 times, and K was determined by the method proposed by Evanno *et al.* (2005). The results of the STRUCTURE HARVESTER analysis indicated that the highest peak of ΔK was at K = 2, while the second, third and fourth peaks of ΔK were also observed at K = 4, K = 6 and K = 8, respectively (data not shown). **Fig. 4** presents sago palm population structures for K = 2, K = 4, K = 6 and K = 8, respectively. These structural analysis results illustrated the possible occurrence of two large sago accession groups in Indonesia, with genetic mixing in each sago palm population from different islands. The accessions of sago palm originating from Sumatra and Kalimantan belonged to a different group than the accessions from Java, Halmahera and Papua. For K = 4, K = 6 or K = 8, it was challenging to determine the number of discrete populations based on **Fig. 4**.

Discussion

This research aimed to generate a partial genome sequence for sago palm and develop SSR markers based on these data. Here we successfully generated a partial sago palm draft genome (315.56 Mb) and demonstrated SSR marker development based on the assembled partial genome. The identified sago palm genome was approximately 13%–18% of the size reported for *Cocos nucifera* (2.42 Gb, Xiao *et al.* 2017), *Elaeis guineensis* (1.8 Gb, Singh *et al.* 2013) and *Arenga pinnata* (1.75 Gb, Rijzaani *et al.* 2017). It was also approximately 50% of the reported *Phoenix dactylifera* genome size (671.2 Mb, Al-Mssallem *et al.* 2013).

Therefore, more data are probably required to obtain the complete sago palm genome. SSR marker development from generated genomes has been reported for various crops (Kale *et al.* 2012, Li *et al.* 2014, Silva *et al.* 2013, Sonah *et al.* 2011, Song *et al.* 2015, Xiao *et al.* 2016, Yang *et al.* 2015). The density of SSR loci (132.57 SSR/Mb) in the identified partial sago palm genome was lower than that reported in other monocot species, which ranged from 175.4 to 363.3 SSR/Mb (Sonah *et al.* 2011). Moreover, the density of SSR loci was also lower than those found in other palm

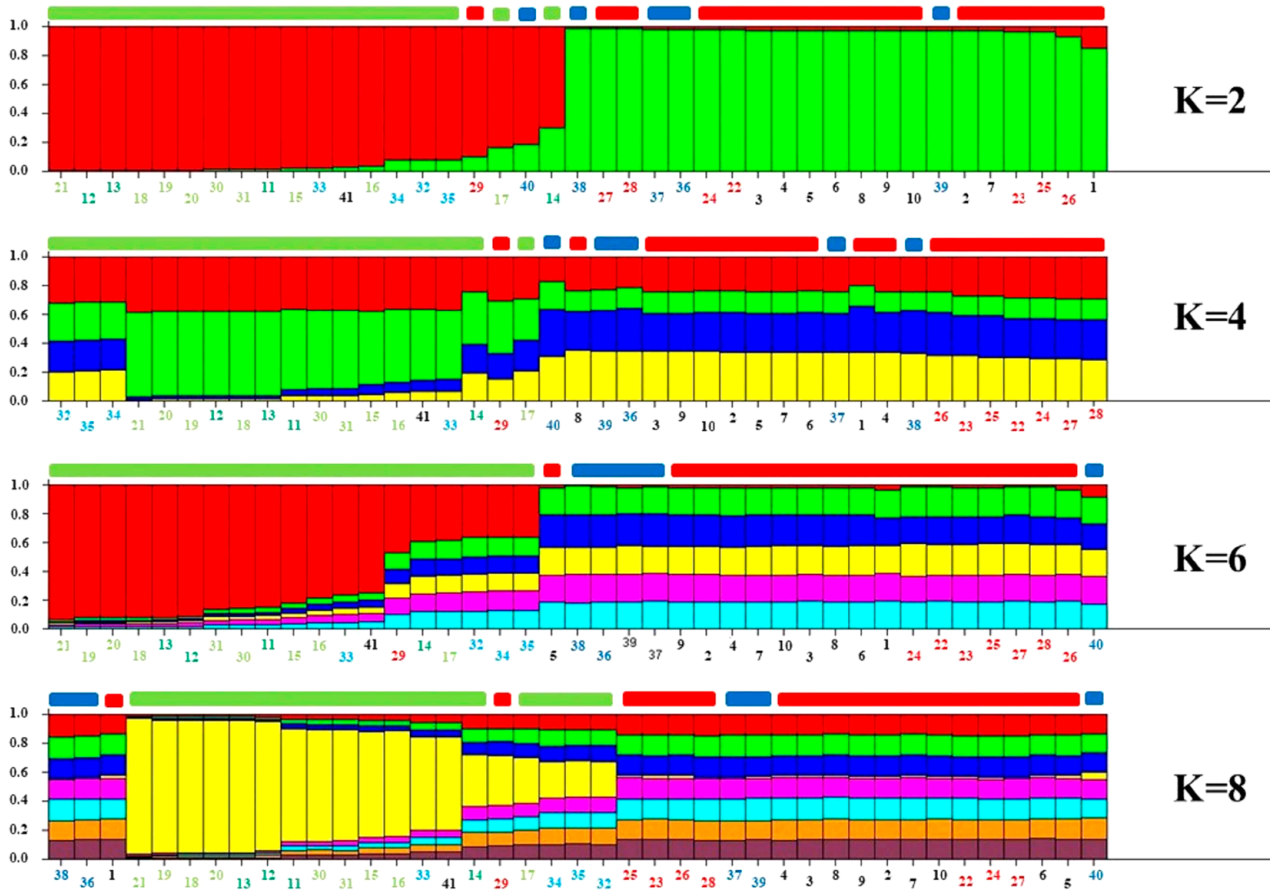

Fig. 4. Population structure of K = 2, K = 4, K = 6 and K = 8 inferred by Bayesian clustering approaches based on seven SSR markers. Samples of sago palm accessions from 1–10, 41: Sumatra Island; 11–14: Halmahera Island; 15–21, 29–31: Java Island; 22–28: Borneo Island; 32–35, 40: Papua Island and 36–39: Seram Island.

species (662.26–696.50 SSR/Mb, Xiao *et al*. 2016). One possible reason for this may be because the sago palm genome sequencing was run only once, which resulted in low resolution NGS data, and the assembled sequences only partially covered the sago palm genome. In the partially identified sago palm genome, AG, AAG and AAAT repeat units were the most frequently found SSR motifs for di-, tri- and tetranucleotide repeats, respectively. Similar results were obtained for oil palm (Ting *et al*. 2010, Zaki *et al*. 2012), date palm (He *et al*. 2017) and wheat (Jaiswal *et al*. 2017).

In this study, we estimated that the densities of penta- and hexanucleotide repeats were at least 10.84 and 9.19 SSR/Mb, respectively. These densities were lower than those found in oil palm, which were estimated as 58.9 SSR/Mb for pentanucleotide and 19.2 SSR/Mb for hexanucleotide repeats (Taeprayoon *et al*. 2015). In their SSR mining, Taeprayoon *et al*. (2015) used a total contig size of 499 Mb from the oil palm genome sequences, whereas we used a total contig size of 238.9 Mb in this sago palm study. In their genome-wide SSR investigation, Xiao *et al*. (2016) suggested that oil palm and date palm have a higher hexanucleotide SSR density than that of some other species.

In Xiao *et al*. (2016) study, they identified a total of 814,383 and 371,629 mono- and hexanucleotide SSRs in the *E. guineensis* and *P. dactylifera* assembled genomic sequences, respectively. They also reported the frequencies of 770.4 and 733.0 SSRs per Mb for mono- and hexanucleotide SSRs (Xiao *et al*. 2016). Tautz and Schlotterer (1994), Klintschar *et al.* (2004), and Song *et al.* (2015) in their studies of two Palmae species found the SSR densities based on the assembled genomic sequences were higher than that in other plant species.

From 500 randomly sampled loci out of 31,659 total SSR loci, 93 (18.6%) SSR primer pairs were identified and synthesised. The validated primers were successfully used to evaluate the genetic diversity of Indonesian sago palm accessions. Our data indicated that the identified SSR primer pairs were more likely to be polymorphic if they were from loci containing dinucleotide SSR motifs. Similar results have also been reported for other plants, indicating that dinucleotides produce more polymorphic alleles than other motifs (Simbaqueba *et al*. 2011,Wang *et al*. 2008).

Using tested and validated SSR primers, we generated 24 different SSR alleles from sago palm, with the number of alleles per locus ranging between two and five across the

sago palm samples. In our study, more alleles were detected in dinucleotide SSRs than in other SSR motifs. The average number of alleles and PIC were 3.429 and 0.475, respectively, indicating that the generated SSR markers could be useful tools for genetic analysis of sago palm germplasm. Based on the criteria developed by Mateescu *et al*. (2005), the average PIC values calculated from seven SSR loci in this study were moderately informative. When tested in sago palm accessions from Indonesia, the SSR markers were more informative than AFLP markers (Kjær *et al*. 2004). Although the calculated PIC of Indonesian sago palm was lower than that reported in date palm (0.67, Arabnezhad *et al*. 2012), it was higher than that of oil palm (0.40, Zaki *et al.* 2012). According to Meszaros *et al*. (2007), molecular markers with moderate and high PIC values were adequate for assessing relationships among accessions based on geographic origin. Compared with AFLP markers previously used to evaluate sago palm, the developed SSR markers should be more useful because SSRs are codominant.

Eleven sago palm samples from Sumatra consisted of nine accessions (B1 to B9) from Bengkulu and two (L and A2) from Lampung provinces. The nine sago palm samples from Bengkulu were taken from provenances close to each other. In nature, vegetative propagation through tillers or suckers results in the formation of sago palm cluster provenances. Because they showed the same genotypes, nine samples taken from Bengkulu province might have been clonal samples. Alternatively, closely related samples from Bengkulu province would have appeared to be genetically identical if the genotyping was done using limited SSR marker loci. Seven SSR marker loci may have been insufficient to differentiate between nine different sago palm samples. However, two sago palm samples (L and A2) from Lampung province were genotyped using the same set of SSR markers and were found to be genetically different compared with the nine samples from Bengkulu province. Therefore, the evaluated SSR marker loci were informative for detecting distantly related sago palm samples from Sumatra. However, further evaluation should be conducted to clarify the clonal status of the nine sago palm samples from Bengkulu using more comprehensive SSR marker loci.

Genetic diversity analysis of Indonesian sago palm populations was previously performed using RAPD markers (Abbas *et al*. 2009), cpDNA (Abbas *et al*. 2010) and waxy gene (Abbas *et al*. 2012). The SSR markers generated in this study offer an alternative approach for evaluating and understanding genetic diversity and determining the relationships among different accessions of sago palms. In comparison with Abbas *et al*. (2009) who evaluated sago palm using RAPD markers, here we evaluated a higher number of sago palm accessions and used SSR markers. However, our results were similar to those obtained by Abbas *et al*. (2009) who also detected at least three different clusters of sago palm. In another study, Abbas *et al*. (2012) reported the presence of two sago palm clusters based on nucleotide variability of the wx gene. Current and previous studies (Abbas *et al*. 2009, 2012) investigated different sago palm accessions, although most of the accessions were from Indonesia. Therefore, it may not be valid to compare one set of results to the others. Moreover, similarities or differences in the findings among these studies may require further validation. Meszaros *et al*. (2007) stated that dissimilarity between groupings obtained using different types of markers might occur because different types of markers access different parts of the genome, even though all selected markers were used to evaluate the same set of organisms.

The results of the current study demonstrated the success of SSR identification and marker development for sago palm based on NGS genome data. The generated SSR markers were used successfully for evaluating the underutilised sago crop genetic diversity. In more comprehensive future studies, additional extensive SSR markers for sago palm based on our current partial genomic resources will be generated. Genotyping more Indonesian sago palm accessions and phenotyping the accessions for various beneficial characters will also be established. By developing additional SSR markers based on partial genome data, genotyping diverse sago palm accessions and phenotyping the studied palm materials, we will provide useful tools for future breeding of this underutilised, carbohydrate-producing crop.

Acknowledgments

The authors wish to thank BPPT for Master's degree scholarship funding between 2014 and 2017 at IPB. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Literature Cited

Abbas, B., M.H. Bintoro, Sudarsono, M. Surahman and H. Ehara (2009) Genetic relationship of sago palm (*Metroxylon sagu* Rottb.) in Indonesia based on RAPD markers. Biodiversitas 10: 168–174.

Abbas, B., Y. Renwarin, M.H. Bintoro, Sudarsono, M. Surahman and H. Ehara (2010) Genetic diversity of sago palm in Indonesia based on chloroplast DNA (cpDNA) markers. Biodiversitas 11: 112–117.

Al-Mssallem, I.S., S. Hu, X. Zhang, Q. Lin, W. Liu, J. Tan, X. Yu, J. Liu, L. Pan, T. Zhang *et al*. (2013) Genome sequence of the date palm *Phoenix dactylifera* L. Nat. Commun. 4: 2274.

Arabnezhad, H., M. Bahar, H.R. Mohammadi and M. Latifian (2012) Development, characterization and use of microsatellite markers for germplasm analysis in date palm (*Phoenix dactylifera* L.). Sci. Hortic. 134: 150–156.

Ashkani, S., M.Y. Rafii, I. Rusli, M. Sariah, S.N.A. Abdullah, H.A. Rahim and M.A. Latif (2012) SSRs for marker-assisted selection for blast resistance in rice (*Oryza sativa* L.). Plant Mol. Biol. Rep. 30: 79–86.

Beccari, O. (1918) Asiatic palms—*Lepidocaryeae*. Annals Royal Botanical Garden, Calcutta. Vol. 12, pp. 156–195.

Brody, J.R. and S.E. Kern (2004) Sodium boric acid: a Tris-free, cooler conductive medium for DNA electrophoresis. BioTechniques 36: 214–216.

Chevallet, M., S. Luche and T. Rabilloud (2006) Silver staining of proteins in polyacrylamide gels. Nat. Protoc. 1: 1852–1858.

Dewi, R.K., M.H. Bintoro and dan Sudradjat (2016) Morphological characteristics and yield potential of sago palm (*Metroxylon* spp.) accessions in South Sorong District, West Papua. J. Agron. Indon. 44: 91–97.

Earl, D.A. and B.M. vonHoldt (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv. Genet. Resour. 4: 359–361.

Evanno, G., S. Regnaut and J. Goudet (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. 14: 2611–2620.

Flach, M. (1997) The Sago Palm. International Plant Genetic Resources Institute, Rome. 73, pp. 8–9.

Geethanjali, S., J. Anitha-Rukmani, D. Rajakumar, P. Kadirvel and P.L. Viswanathan (2018) Genetic diversity, population structure and association analysis in coconut (*Cocos nucifera* L.) germplasm using SSR markers. Plant Genet. Resour. 16: 156–168.

Girichev, V., M.V. Hanke, A. Peil and H. Flachowsky (2017) SSR fingerprinting of a German *Rubus* collection and pedigree based evaluation on trueness-to-type. Genet. Resour. Crop Evol. 64: 189–203.

Grover, A., V. Aishwarya and P.C. Sharma (2012) Searching microsatellites in DNA sequences: approaches used and tools developed. Physiol. Mol. Biol. Plants 18: 11–19.

Guo, W.J., J. Ling and P. Li (2009) Consensus features of microsatellite distribution: Microsatellite contents are universally correlated with recombination rates and are preferentially depressed by centromeres in multicellular eukaryotic genomes. Genomics 93: 323–331.

He, Z., C. Zhang, W. Liu, Q. Lin, T. Wei, H.A. Aljohi, W.H. Chen and S. Hu (2017) DRDB: An online date palm genomic resource Database. Front. Plant Sci. 8: 1889.

Jaiswal, S., S. Sheoran, V. Arora, U.B. Angadi, M.A. Iquebal, N. Raghav, B. Aneja, D. Kumar, R. Singh, P. Sharma *et al.* (2017) Putative microsatellite DNA marker-based wheat genomic resource for varietal improvement and management. Front. Plant Sci. 8: 2009.

Kale, S.M., V.C. Pardeshi, N.Y. Kadoo, P.B. Ghorpade, M.M. Jana and V.S. Gupta (2012) Development of genomic simple sequence repeat markers for linseed using next-generation sequencing technology. Mol. Breed. 30: 597–606.

Kalinowski, S.T., M.L. Taper and T.C. Marshall (2007) Revising how the computer program Cervus accommodates genotyping error increases success in paternity assignment. Mol. Ecol. 16: 1099–1106.

Karim, A.A., A.P. Tie, D.M.A. Manan and I.S.M. Zaidul (2008) Starch from the sago (*Metroxylon sagu*) palm tree—properties, prospects, and challenges as a new industrial source for food and other uses. Compr. Rev. Food Sci. Food Saf. 7: 215–228.

Kaur, K., V. Sharma, V. Singh, M.S. Wani and R.C. Gupta (2016) Development of novel SSR markers for evaluation of genetic diversity and population structure in *Tribulus terrestris* L. (*Zygophyllaceae*). 3 Biotech 6: 156.

Kelkar, Y.D., S. Tyekucheva, F. Chiaromonte and M.D. Makova (2008) The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res. 18: 30–38.

Kjær, A., A.S. Barfod, C.B. Asmussen and O. Seberg (2004) Investigation of genetic and morphological variation in the sago palm (*Metroxylon sagu*; Arecaceae) in Papua New Guinea. Ann. Bot. 94: 109–117.

Li, W., Y. Feng, H. Sun, Y. Deng, H. Yu and H. Chen (2014) Analysis of simple sequence repeats in the *Gaeumannomyces graminis* var. *tritici* genome and the development of microsatellite markers. Curr. Genet. 60: 237–245.

Maskromo, I., S.H. Larekeng, H. Novarianto and S. Sudarsono (2016) Xenia negatively affecting kopyor nut yield in Kalianda Tall Kopyor and Pati Dwarf Kopyor Coconuts. Emir. J. Food Agric. 28: 644–652.

Mateescu, R.G., Z. Zhang, K. Tsai, J. Phavaphutanon, N.I. Burton-Wurster, G. Lust, R. Quaas, K. Murphy, G.M. Acland and R.J. Todhunter (2005) Analysis of allele fidelity, polymorphic information content and density of microsatellites in a genome-wide screening for hip dysplasia in a crossbreed pedigree. J. Hered. 96: 847–853.

Meszaros, K., I. Karsai, C. Kuti, J. Banyai, L. Lang and Z. Bedo (2007) Efficiency of different marker systems for genotype fingerprinting and genetic diversity studies in barley (*Hordeum vulgare* L.) S. Afr. J. Bot. 73: 43–48.

Miyamoto, E., S. Matsuda, H. Ando, K. Kakuda, F.S. Jong and A. Watanabe (2009) Effect of sago palm (*Metroxylon sagu* Rottb.) cultivation on the chemical properties of soil and water in tropical peat soil ecosystem. Nutr. Cycl. Agroecosyst. 85: 157–167.

Novero, A.U., B.M. Ma and J.E. Hannah (2012) Epigenetic inheritance of spine formation in sago palm (*Metroxylon sagu* Roettb.). Plant Omics J. 5: 559–566.

Ott, A., B. Trautschold and D. Sandhu (2011) Using microsatellites to understand the physical distribution of recombination on soybean chromosomes. PLoS ONE 6: e22306.

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

Perrier, X. and J.P. Jacquemoud-Collet (2006) DARWin software. http://darwin.cirad.fr/darwin.

Pesik, A., D. Efendi, H. Novarianto, D. Dinarti, I. Maskromo, E.T. Tenda and Sudarsono (2015) Keragaman dan Hubungan Genetik Antara Kelapa Tetua Genjah Kuning Nias Bul. Palma 16: 129–140.

Pesik, A., D. Efendi, H. Novarianto, D. Dinarti and Sudarsono (2017) Development of SNAP markers based on nucleotide variability of WRKY genes in coconut and their validation using multiplex PCR. Biodiversitas 18: 465–475.

Pritchard, J.K., M. Stephens and P. Donnelly (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

Rauscher, G. and I. Simko (2013) Development of genomic SSR markers for fingerprinting lettuce (*Lactuca sativa* L.) cultivars and mapping genes. BMC Plant Biol. 13: 11.

Rijzaani, H., P. Lestari, I. Maskromo, Surdarsono and T.P. Priyatno (2017) Assembly of Sugar Palm (*Arenga pinnata* Wurmb Merr.) draft genome. Warta Penelitian Pengembangan Pertanian. 39: 10–12.

Sharma, P.C., A. Grover and G. Kahl (2007) Mining microsatellites in eukaryotic genomes. Trends Biotechnol. 25: 490–498.

Silva, P.I.T., A.M. Martins, E.G. Gouvea, M. Pessoa-Filho and M.E. Ferreira (2013) Development and validation of microsatellite markers for *Brachiaria ruziziensis* obtained by partial genome assembly of Illumina single-end reads. BMC Genomics 14: 17.

Singh Kesawat, M. and B.K. Das (2009) Molecular markers: Its application in crop improvement. J. Crop Sci. Biotechnol. 12: 169–181.

Singh, R., M. Ong-Abdullah, E.L. Low, M.A. Manaf, R. Rosli, R. Nookiah, L.C. Li Ooi, S. Eng Ooi, K.L. Chan, M.A. Halim *et al.* (2013) Oil palm genome sequence reveals divergence of interfertile species in old and new worlds. Nature 500: 335–339.

Singhal, R.S., J.F. Kennedy, S.M. Gopalakrishnan, A. Kaczmarek, C.J. Knill and P.F. Akmar (2008) Industrial production, processing, and utilization of sago palm-derived products. Carbohyd. Polym. 72:

1–20.

Sonah, H., R.K. Deshmukh, A. Sharma, V.P. Singh, D.K. Gupta, R.N. Gacche, J.C. Rana, N.K. Singh and T.R. Sharma (2011) Genome-wide distribution and organization of microsatellites in plants: an insight into marker development in *Brachypodium*. PLoS ONE 6: e21298.

Song, X., T. Ge, Y. Li and X. Hou (2015) Genome-wide identification of SSR and SNP markers from the non-heading Chinese cabbage for comparative genomic analysis. BMC Genomics 16: 328.

Taeprayoon, P., T. Tanya, Y.J. Kang, A. Limsrivilai, S.H. Lee and P. Srinives (2015) Genome-wide SSR marker development in oil palm by Illumina HiSeq for parental selection. Plant Genet. Resour. 14: 157–160.

Tajuddin, T., K. Karyanti, M. Minaldi and N. Haska (2008) Sago: its Potential in Food and Industry Proceedings of the 9th International Sago Symposium. Visayas State University, TUAT Press, The Philippines, pp. 231–235.

Tinche, D. Asmono, D. Dinarty and Sudarsono (2014) Genetic diversity oil palm (*Elaeis guineensis* Jacq.) Nigeria population based on SSR (Simple Sequence Repeats) marker analysis. Bul. Palma 15: 14–23.

Ting, N.C., N.M. Zaki, R. Rosli, E.-T. Low, M. Ithnin, S.-C. Cheah, S.-G. Tan and R. Singh (2010) SSR mining in oil palm EST database: application in oil palm germplasm diversity studies. J. Genet. 89: 135–145.

Uthumporn, U., N. Wahidah and A.A. Karim (2014) Physicochemical properties of starch from sago (*Metroxylon Sagu*) palm grown in mineral soil at different growth stages. IOP Conf. Ser. Mater. Sci. Eng. 62: 247.

Wang, J., C. Chen, J. Na, Q. Yu, S. Hou, R.E. Paull, P.H. Moore, M. Alam and R. Ming (2008) Genome-wide comparative analyses of microsatellites in papaya. Trop. Plant Biol. 1: 278–292.

Wee, C.C. and H.A. Roslan (2012) Expressed sequence tags (ESTs) from young leaves of *Metroxylon sagu*. 3 Biotech 2: 211–218.

Xiao, Y., W. Xia, J. Ma, A.S. Mason, H. Fan, P. Shi, X. Lei, Z. Ma and M. Peng (2016) Genome-wide identification and transferability of microsatellite markers between Palmae species. Front. Plant Sci. 7: 1578.

Xiao, Y., P. Xu, H. Fan, L. Baudouin, W. Xia, S. Bocs, J. Xu, Q. Li, A. Guo, L. Zhou *et al.* (2017) The genome draft of coconut (*Cocos nucifera*). Gigascience 6: gix095.

Yang, T., L. Fang, X. Zhang, J. Hu, S. Bao, J. Hao, L. Li, Y. He, J. Jiang, F. Wang *et al.* (2015) High-throughput development of SSR markers from pea (*Pisum sativum* L.) based on next-generation sequencing of a purified Chinese commercial variety. PLoS ONE 10: e0139775.

Zaki, N.M., R. Singh, R. Rosli and I. Ismail (2012) *Elaeis oleifera* genomic-SSR markers: Exploitation in oil palm germplasm diversity and cross-amplification in Arecaceae. Int. J. Mol. Sci. 13: 4069–4088.

Zalapa, J.E., H. Cuevas, H. Zhu, S. Steffan, D. Senalik, E. Zeldin, B. Mccown, R. Harbut and P. Simon (2012) Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. Am. J. Bot. 99: 193–208.

Zong, J.W., T.T. Zhao, Q.H. Ma, L.S. Liang and G.X. Wang (2015) Assessment of genetic diversity and population genetic structure of *Corylus mandshurica* in China using SSR markers. PLoS ONE 10: e0137528.