Mining and characterization of genomic-based microsatellite markers in duku (*Lansium domesticum*)

H P Sari¹, D Efendi², W B Suwarno² and D D Matra²*

¹Study Program of Plant Breeding and Biotechnology, Graduate School of IPB University
²Department of Agronomy and Horticulture, IPB University, Meranti road, IPB Dramaga Campus Bogor-West Java-Indonesia 16680

*Correspondence author: dedenmatra@apps.ipb.ac.id

Abstract. Duku (*Lansium domesticum*) has a high diversity. There are no specific markers to assess genetic diversity in Duku. Microsatellite markers can determine genetic diversity more specifically in the cultivar level. This study aimed to isolate and characterize microsatellite sequences from the assembled-genome database. The markers were developed from 455 010 contigs of the velvet-assembled genome using the MISA program. BLASTN program was used for sequence annotation. Characterization of microsatellite markers was analyzed using bioinformatics. The result was found 7 types of repeat type: 63 contigs of mononucleotide with most motif of T, 450 contigs of dinucleotide with most motif of TA, 3094 contigs of trinucleotide with most motif of AAT, 561 contigs of tetranucleotide with most motif of TTAA, 85 contigs of pentanucleotide with most motif of TTTCT, 144 contigs of hexanucleotide with most motif of AAAAAT and 1331 contigs of compound sequence. This study provides information about specific markers in *L. domesticum* and will contribute to plant breeding development.

Key words: De novo genome, genetic diversity, phylogenetic, contigs, plant breeding

1. Introduction

Duku (*Lansium domesticum*) originating from Southeast Asia such as Indonesia, the Philippines, Malaysia and Cambodia [1]. Fruit that has pale yellow skin is popular because it tastes sweet and refreshing. Besides being consumed as fresh fruit, Lim *et al* [2] showed that this fruit has a high antioxidant content.

Indonesia has various types of *L. domesticum* which are widespread on the islands of Sumatra, Kalimantan and Java. This species has a high diversity morphologically [3]. Study of genetic diversity through morphology has several obstacles to do especially on annual plants because it requires a long time and some characters cannot be observed directly. Another obstacle is that selection using morphological characters makes the desired traits and undesirable traits difficult to separate when crossing [4]. Another alternative to overcome these obstacles is to use molecular markers.

Molecular markers are DNA sequences that can be identified and inherited and are present in certain locations in the genome and characterize certain characteristics. Zulfahmi [5] states that molecular markers are more stable in terms of observation compared to morphological markers and
can be detected in all plant tissues, and are not influenced by the environment. This analysis also generally gives more accurate results compared to morphological analysis.

There are two categories of markers commonly used in molecular analysis, namely: first, DNA markers without PCR (Polymerase Chain Reaction) such as RFLP, second, DNA markers based on PCR which include RAPD, AFLP, SSR, CAPS, SCAR, SSCP and DNA Barcoding [5]. For diversity analysis, people used to use RAPD (Random Amplified Polymorphic DNA) markers. Longkong (kokosan) has the highest DNA content while duku has the lowest DNA and langsat was intermediate based on RAPD markers.

As knowledge develops, RAPD is judged to be less specific in determining polymorphic DNA so that microsatellite markers (1-10 nucleotides) are chosen alternatives. Microsatellite markers are known to have very high polymorphic properties. In addition, Microsatellite markers are codominant, the quantity of DNA needed is not much, the method used is quite simple and widely available in the market [6]. Microsatellite markers are able to show generate more codominant character and high levels of allelic variation in lychee plants information per unit test than other markers. The same thing was reported by Napitu et al. [7] in wild rambutan species. Aljumaili et al. [8] states that Microsatellite is a technique suitable for the study of genetic diversity and is capable of producing high diversity grouping patterns. marker can determine genetic diversity more specifically in the cultivar level.

Molecular markers can be widely developed through the application of NGS (Next Generation Sequencing). Matra et al. [9] had sequenced the L. domesticum species genome by whole genome sequencing technique. This study aims to isolate and characterize microsatellite sequences from the assembled-genome database.

2. Materials and methods

This research was conducted in July-August 2019 at the Department of Agronomy and Horticulture IPB University. The data used is the result of sequencing using Next Generation Sequencing (NGS) conducted by Matra et al. [9]. The Assembly-stat program is used to calculate contig statistics. The contigs having minimum filtering of 200 bp and filtering from the redundancy contig using the CAP3 and CD-hit programs. MISA program is used to identify content containing microsatellite with minimum repetitions: 10 for 1 basis, 6 for 2 bases, and 5 for 3, 4, 5, and 6 bases; and interruptions (the maximum difference between microsatellites) are 100 bases. The research mechanism can be seen in figure 1.

3. Result and discussion

The data used is the result of genome DNA sequencing. 5728 microsatellites were identified in this study. After functional analysis, several types of nucleotides have been obtained, including: mono-, di-, tri-, tetra-, penta-, and hexa-nucleotide (figure 2). Mononucleotides are generally not used for primary designs because they are considered less informative. So that what is commonly used is hexanucleotide. In living things, the most common types of nucleotides are dinucleotides and trinucleotides. The result was found 7 types of repeat types: 63 contigs of mononucleotide, 450 contigs of dinucleotide, 3094 contigs of trinucleotide, 561 contigs of tetranucleotide, 85 contigs of penta nucleotide, 144 contigs of hexanucleotide and 1331 contigs of compound sequence.

Microsatellites can be categorized based on their motifs with the composition: i) perfect, if the whole consists of repetition of single motive; ii) imperfect, if base pairs are not included in the motive to occur between repetitions; iii) interrupted if the order of some base pairs is included in the motif; or iv) composite if formed by many, close together, repetitive motifs [10-12]. In this study, almost all of them were included in perfect order and combined or combined. Specific primary synthesis can be carried out after identifying the sequence containing the SSR. Specific primers must be synthesized with a length of 18 to 25 bp, complementary to the flanking area, and followed by an amplification and polymorphism test.
Figure 1. Pipeline for microsatellite marker development using NGS data.

Figure 2. Percentage of DNA contigs containing microsatellite.

The major motifs were rich in T, TA, AAT, TTAA, TTTCT, AAAAAT and the minor motifs were mostly rich in G / C. The distributions of microsatellites types in assembled genomic sequences of L. domesticum are presented in figure 3.
The total DNA contig from NGS data is 5728 contig. Bioinformatic searches of 1-6 microsatellite nucleotides indicate that 54% contain the AAT sequence, this sequence is the most abundant sequence in the genome. This is similar to the study of Aibin et al. [13] who reported that microsatellite sequences in abalone showed abundant trinucleotide which was 47%. According to Powel et al. [14] the most frequent repetition motives in plants are the AT, AG, and TC motives. However this is quite different from the results obtained in L. domesticum species. The distribution of SSR motifs in the L. domesticum genome is dominated by the type of tri-nucleotide with the AAT motif. It is natural in the opinion of Li et al. [15] that the SSR motives for each species vary.

Figure 3. Major motifs

4. Conclusion
This study provides information about microsatellite markers of L. domesticum. The primary design for molecular analysis can be carried out based on these sequences. The results of molecular analysis are expected to contribute to the development of plant breeding either as DNA-Barcoding, development of new varieties or other purposes.

Acknowledgements
Thank you for the support of the INSINAS Program (Research Incentives Program for the National Innovation System) on the “Pratama Research “ scheme (12/INS-1/PPK/E4/2018) from the Directorate General of Research and Development Strengthening Ministry of Research, Technology and Higher Education Republic of Indonesia.

References
[1] Orwa C, A Mutua, Kindt, Jamnadas R, and Anthony S 2009 Agroforestree Database: a Tree Reference and Selection Guide Version 4.0. (Nairobi: World Agroforestry Centre)
[2] Lim Y Y, Lim T T and Tee J 2007 Science Direct (Food Chemistry) 103 1003-1008
[3] Yulita K S 2011 Biodiversitas 12 125-130
[4] Azrai M 2005 J. Agrobiogen 1 26-37
[5] Zulfahmi 2013 J. Agroteknologi 3 41-52
[6] Madhou M, Normand F, Bahorun T and Hormaza J J 2013 Tree Genet and Gen 9 387-396
[7] Napitu C S, Djuita N R dan Chikmawati T 2016 Floribunda 5 4
[8] Aljumaili S J, Rafii M Y, Latif M A, Sakimin S Z, Arolu I W and Miah G 2018 BioMed Research International 11
[9] Matra D D, Ritonga A W, Natawijaya A, Poerwanto R, Sobir, Siregar U J, Widodo W D, Eiichi E D 2019 Data in Brief 22 960–963
[10] Oliveira E J, Padua J G, Zucchi M I, Vencovsky R, and Vieira M L C 2006 Genet Mol Biol 29 294-307
[11] Mason A S 2015 SSR Genotyping In: Batley Springer 77- 89
[12] Vieira M L C, Santini L, Diniz A L and Munhoz C F 2016 Genetics and Molecular Biology 39 312-328
[13] Aibin Z, Zhenmin B, Mingling W, Dan C, Jian Y, Xiaolong W, Xiaoli H, Chengzhu, and Jingjie H 2008 J. Ocean Univ. Chin. (Oceanic and Coastal Sea Research) 7 219-222
[14] Powel W, Gordon C M and Jim P 1996 Trends Plant Sci 1 215-221
[15] Li Y C, Korol A B, Fahima T, Beiles A, and Nevo E  2002 Molecular Ecology 11 2453-2465