Simple sequence repeats (SSRs) discovery and characterization from *Phoenix dactylifera* genome

Aditya Nugroho[1], *Muhammad Roiyan* Romadhon[2*], and *Efratenta Katherina* Depari[3]

[1]Department of Silviculture, Faculty of Forestry and Environment, IPB University, Bogor, West Java, Indonesia
[2]Center for Standard Testing of Palm Plant Instruments (BSIP Palma), Manado, North Sulawesi, Indonesia
[3]Department of Forestry, Faculty of Agriculture, Bengkulu University, Bengkulu, Bengkulu, Indonesia

Abstract. The date palm is a palm plant belonging to the Arecaceae family. Date palms have several benefits, such as leaves used in various religious ceremonies, tree trunks as firewood, and fruit with various health benefits. In addition, date palm flowers undergo cross-pollination, are dioecious, and consist of 18 chromosomes, resulting in a heterozygous genetic constitution that can lead to high genetic diversity. The development of Next Generation Sequencing technology can detect genetic diversity using whole genome sequencing approaches. Therefore, this study aims to discover and develop SSR markers using an in-silico approach from Phoenix dactylifera genome data. Genome data from male Phoenix dactylifera leaf tissue was obtained from NCBI with GenBank assembly accession: GCA_009389715.1. Quality analysis of de novo assembly using Busco Analysis result in single-copy completeness of 76.7%, duplicated completeness of 19.0%, fragmented completeness of 2.8%, and missing completeness of 1.5%. A total of 36,764 genes and 29,239 protein-coding genes were found. SSRs were identified and extracted using the Microsatellite (MISA) program, resulting in the distribution of dinucleotide SSR motifs (68.37%), trinucleotide (24.22%), tetranucleotide (6.36%), pentanucleotide (0.59%), and heptanucleotide (0.47%). Based on these perfect SSRs, 15 primer pairs were designed. The SSR markers developed will be expected to help further research on the genetic diversity of *P. dactylifera.*

*Corresponding author: ~~mroiyanripb1@gmail.com~~

1 Introduction

Dates (*Phoenix dactylifera*) are a member of the Palmae family and are dioecious, which means that female and male blooms grow on distinct trees. The date commodity in Indonesia has developed and become one of the leading plantation commodities in East Java. Date plants have a high degree of genetic diversity due to cross-pollination. Evaluation activities related to the genetic diversity of dates and the estimation of genetic parameters are necessary to predict the outcomes of selection and the selection strategies used for the germplasm collection, with the main goal of developing superior date hybrids [1-2].

A plant genetic information is highly beneficial for the breeding of perennial plants. Genetic information on palm plants, particularly date palms, still very limited. According to [3], conservation activities for perennial plants are carried out based on scientific data through molecular analysis and whole genome sequencing. Genome sequencing activities are very easy and fast, thus aiding in the genetic improvement and innovation of palm plants. NGS (Next-Generation Sequencing) is a DNA sequencing technology developed after the Sanger method, designed to perform DNA sequencing in parallel and on a large scale. NGS allowed researchers to sequence millions to billions of DNA nucleotides in a single run. NGS has been used in various applications, including is whole genome sequencing (WGS), which is very useful in producing long, high-quality sequence reads.

The genetic information obtained from DNA sequencing of date palms is highly beneficial for research in functional gene studies, genetic diversity, and the construction of phylogenetic relationships. The genetic improvement of areca nut plants involve DNA modifications, beginning with DNA isolation activities, and requires genomic information. Genomic information related to areca nut can be obtained through WGS of dates. WGS has been extensively performed with bioinformatics analysis on food crops such as soybeans [4-6], rice [7-10], and maize [11]. WGS has also been conducted on plantation crops such as oil palm [12], coconut [13], and areca nut [14].

The molecular approach using SSR (Single Sequence Repeat) markers are one method for identifying the genetic diversity of dates. Isolating SSR from WGS data is a solution to obtain specific primers for date palm diversity. SSR markers are co-dominant markers that can distinguish heterozygotes from date palm alleles, making it a rapid approach for assessing the genetic diversity of date palms in Indonesia. Therefore, this research aims to discover and develop SSR markers using an in-silico approach.

2 Methods

The materials used in this study included the genome sequences of *P. dactylifera*, which have been deposited in the NCBI database with BioSample ID: SAMN05011615 from a previous study [15]. Simple Sequence Repeats were detected and extracted using the Microsatellite (MISA) program. Minimum repeats were determined using the following criteria: 10 repeats for dinucleotide motifs and six repeats for tri-, tetra-, penta-, and hexanucleotide motifs. Additionally, interruptions, defined as the maximum allowable distance between adjacent microsatellites, were set to 100 bases.

Marker development was carried out using the Primer3Plus tool (https://www.primer3plus.com). The primer design parameters included a minimum and maximum amplicon size of 100–300 base pairs (bp). The primer length parameters were set with a minimum of 18 bp, an optimum of 20 bp, and a maximum of 23 bp. The melting temperature (Tm) range for primers was specified to be between 59–62 °C, and the GC content of the primers were between 40–80%.

3 Results

In this study, a total of 18 genomic sequences were examined, encompassing a combined length of 385,590,432 base pairs (bp). From this analysis, we identified 31,594 SSRs distributed across the examined sequences (Table 1). Interestingly, all 18 genomic sequences contained SSRs, and each sequence contained more than one SSR, indicating a wide and diverse distribution of SSR motifs within the *P. dactylifera* genome. Specifically, 4,068 of the identified SSRs were found in compound formations, suggesting the presence of more complex patterns in the arrangement of these motifs.

Table 1 Statistics of microsatellite discovery

Parameter	**Count**
Total number of sequences examined	18
Total size of examined sequences (bp)	385,590,432
Total number of identified SSRs	31,594
Number of SSR-containing sequences	18
Number of sequences containing more than 1 SSR	18
Number of SSRs present in compound formation	4,068

Compound SSRs (cSSRs), are a type of simple sequence repeat (SSR) motif consisting of two or more microsatellites. The motifs consist of different SSR combinations, with perfect or imperfect repetition, so they cannot be used to identify diversity. The results of SSR discovery in the genome of *P. dactylifera* show that dinucleotide SSR motifs are the most dominant type, comprising 68% of all identified SSRs. Trinucleotide motifs rank second with a proportion of 24%, followed by tetranucleotide motifs at 6%. Meanwhile, pentanucleotide and hexanucleotide motifs each account for only 1% of the total SSRs found (Figure 1). This distribution showed a common pattern in many genomes where shorter SSR motifs are be more prevalent than longer ones.

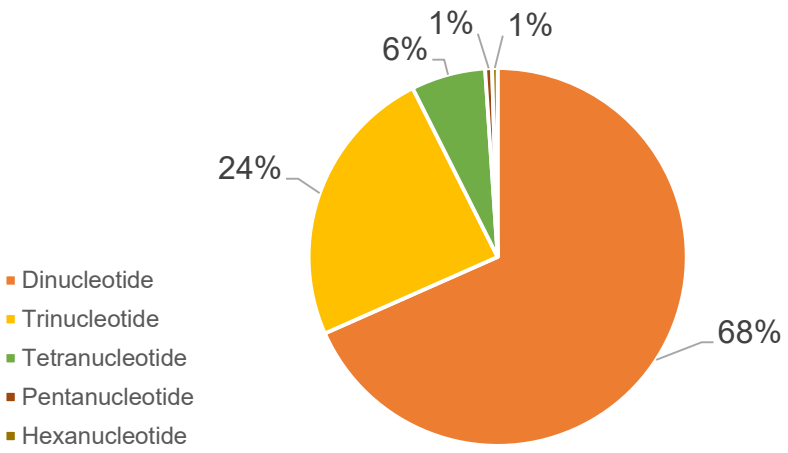

Figure 1 Distribution of SSR motifs

The highest frequency of SSR motifs found in the *P. dactylifera* genome is the "AT" motif, with 5672 occurrences, followed by the "TA" motif, with 3944 occurrences. Meanwhile, the lowest frequencies in this data are the "TT" and "CC" motifs, with frequencies of 180 and 131 respectively (Figure 2). These discovered motifs were then developed into SSR markers. The complete primers generated from this research could be accessed at https://data.mendeley.com/datasets/wmvrj6x5yb/1.

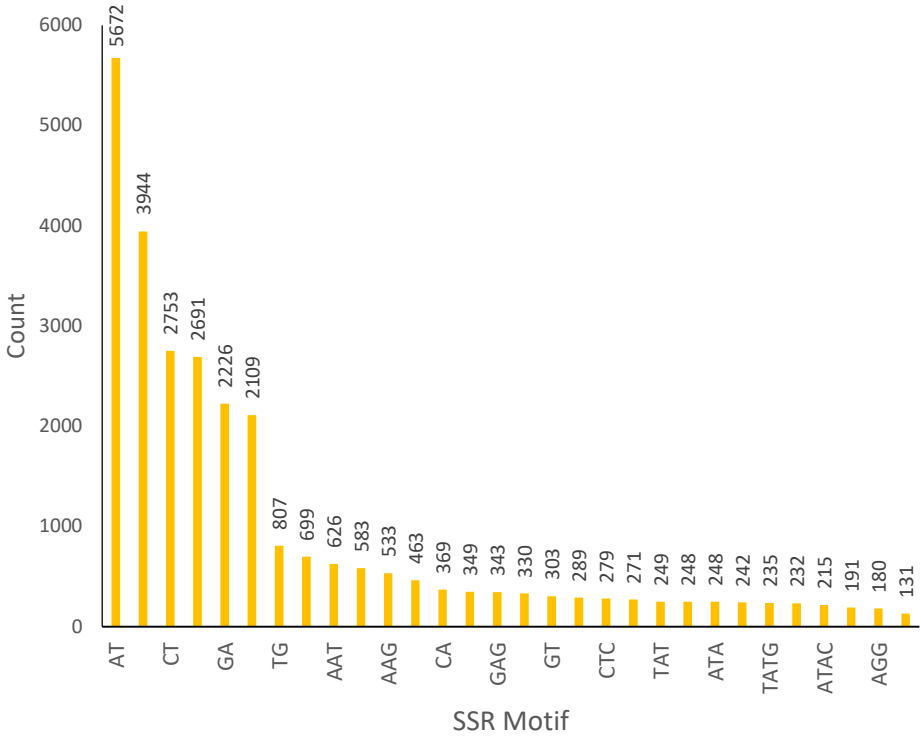

Figure 2 Frequency of SSR motif identified

4 Discussion

This study revealed that *Phoenix dactylifera* has a wide and diverse distribution of SSRs in its genome, with 31,594 SSRs identified from the whole genome analyzed. The even distribution of SSRs within each sequence suggests that SSRs play a major role in the structure and function of the *P. dactylifera* genome. This widespread presence and variability of SSRs is consistent with findings in other plant genomes, where SSRs were known to play an important role in genome evolution and genetic diversity [16-18].

The distribution of dinucleotide motifs that reached 68% dominated all motifs identified in the genome (Figure 1). This result was consistent with [19] reported that dinucleotide motifs dominated the SSR identification results in the genome and transcriptome of areca nut. Previously [20] also found dinucleotide motifs to have the highest abundance in the genome of five species in the Palmae family: *P. dactylifera*, *C. nucifera*, *C. simplicifolius*, *E. oleifera*, and *E. guineensis*. Dinucleotide motifs were most commonly found in SSR identification resulted because these motifs have very stable properties and were easily found in the genome [21-22]. [23] observed that SSR motifs and their variations in maize

follow a specific pattern, decreasing from dinucleotide to hexanucleotide motifs. The pattern was similar to the results of this study, where pentanucleotide and hexanucleotide motifs showed the most minor distribution (1%).

The high frequency of AT and TA motifs in SSR is closely related to the presence of these motifs in non-coding regions, intergenic spaces, and introns [23]. These motifs are often found in regions of the genome that were not transcribed into proteins, which indicates that they may play a role in genome structure and function without being directly involved in protein synthesis. The stability of AT-rich motifs across plant developmental phases provides strong evidence for the importance of these motifs in different stages of plant life. The tolerance of AT motifs to environmental influences suggests that these motifs might confer a selective advantage on the plants that contain them, allowing the plants to adapt and survive in a wide range of environmental conditions [24].

SSR is frequently utilized as a marker in functional genomics studies, population genetics, association mapping, DNA fingerprinting, diversity analysis, comparative mapping, and gene tagging due to its excellent reproducibility, a high level of polymorphism, and high mutation rate. SSR markers are co-dominant, providing information on whether the targeted locus was heterozygous or homozygous [25]. SSR has been widely used for genetic diversity analysis in many palm families, i.e, oil palm, sugar palm, areca nut, coconut, and nipa [26-30]. These studies indicate that the markers established in this research are promising for analysing genetic diversity in date palms.

5 Conclusion

SSR characterization in this study provided insight into the abundance and distribution in the genome of *P. dactylifera*. A total of 31 594 SSRs were identified in the *P. dactylifera*. The most abundant type of SSRs was found in dinucleotide (68.37%), followed by trinucleotide (24.22%). AT (5672) and TA (3944) were the most common SSR motifs. SSR markers produced in this study could be used as an alternative primer in the study of the genetic diversity of *P. dactylifera*.

References

1. M. Fisher, Moving science forward through: Meta-analysis, (CSA News, 2015)
2. H.D. Toler, R.M. Augé, V. Benelli, F.L. Allen, A.J. Ashworth, Global meta-analysis of cotton yield and weed suppression from cover crops, CSSA. **59**, 1248-1260 (2019)
3. Soenarno, S. Astana, Lacak balak untuk verifikasi uji legalitas kayu pada pemanenan kayu hutan alam, JPHT. **36**, 47-58 (2018)
4. D. Kovalic, C. Garnaa, L. Guo, Y. Yan, J. Groat, A. Silvanovich, The use of next generation sequencing and junction sequence analysis bioinformatics to achieve molecular characterization of crops improved through modern biotechnology, Plant Genome. **5**, 149–163 (2022)
5. B. Guo, Y. Guo, H. Hong, L.J. Qiu, Identification of genomic insertion and flanking sequence of G2-EPSPS and GAT transgenes in soybean using whole genome sequencing method, Front. Plant Sci. **7**, 1009 (2016)
6. S.K. Guttikonda, P. Marri, J. Mammadov, L. Ye, K. Soe, K. Richey, Molecular characterization of transgenic events using next generation sequencing approach, PLoS One. **11**, e0149515 (2016)

7. L. Yang, C. Wang, A. Holst-Jensen, D. Morisset, Y. Lin, D. Zhang, Characterization of GM events by insert knowledge adapted re-sequencing approaches, Sci. Rep. **3**, 2839 (2013)
8. D. Park, D. Kim, G. Jang, J. Lim, Y.J. Shin, J. Kim, Efficiency to discovery transgenic loci in GM rice using next generation sequencing whole genome re-sequencing, Genom. Inform. **13**, 81–85 (2015)
9. D. Park, S.H. Park, Y.W. Ban, Y.S. Kim, K.C. Park, N.S. Kim, A bioinformatics approach for identifying transgene insertion sites using whole genome sequencing data, BMC Biotechnol. **17**, 67 (2017)
10. Y. Zhang, H. Zhang, Z. Qu, X. Zhang, J. Cu, C. Wang, Comprehensive analysis of the molecular characterization of GM rice G6H1 using a paired-end sequencing approach, Food Chem. **309**, 125760 (2020)
11. R. Cade, K. Burgin, K. Schilling, T.J. Lee, P. Ngam, N. Devitt, Evaluation of whole genome sequencing and an insertion site characterization method for molecular characterization of GM maize, J. Regul. Sci. **6**, 1–14 (2018)
12. J. Jin, M. Lee, B. Bai, Y. Sun, J. Qu, Rahmadsyah, Y. Alfiko, C.H. Lim, A. Suwanto, M. Sugiharti, L. Wong, J. Ye, N.H. Chua, G.H. Yue, Draft genome sequence of an elite Dura palm and whole-genome patterns of DNA variation in oil palm, Dna Res. **23**, 527–533 (2016)
13. Y. Yang, S. Bocs, H. Fan, Coconut genome assembly enables evolutionary analysis of palms and highlights signaling pathways involved in salt tolerance, Commun. Biol. **4**, 1-14 (2021)
14. Y. Yang, L. Huang, C. Xu, L. Qi, Z. Wu, J. Li, H. Chen, Y. Wu, T. Fu, H. Zhu, M.A. Saand, J. Li, L. Liu, H. Fan, H. Zhou, W. Qin, Chromosome-scale genome assembly of areca palm (*Areca catechu*), Mol Ecol. Resour. **21**, 1–16 (2021)
15. K.M. Hazzouri, M. Gros-Balthazard, J.M. Flowers, D. Copetti, A. Lemansour, M. Lebrun, M.D. Purugganan, Genome-wide association mapping of date palm fruit traits, Nat. Commun. **10**, 4680 (2019)
16. C. Liu, J. Li, G. Qin, Genome-wide distribution of simple sequence repeats in pomegranate and their application to the analysis of genetic diversity, Tree Genet. Genomes. **16**, 1-9 (2020)
17. J. Ping, P. Feng, J. Li, R. Zhang, Y. Su, T. Wang, Molecular evolution and SSRs analysis based on the chloroplast genome of *Callitropsis funebris*, Ecol. Evol. **11**, 4786-4802 (2021)
18. M. Zhao, G. Shu, Y. Hu, G. Cao, Y. Wang, Pattern and variation in simple sequence repeat (SSR) at different genomic regions and its implications to maize evolution and breeding, BMC Genomics. **24**, 136 (2023)
19. M.R. Romadhon, S. Sobir, W.B. Suwarno, D.D. Matra, Profile microsatellite mining of whole genome sequencing and transcriptomic assembly in dwarf and tall areca nut (*Areca catechu*) in Indonesia, Biodiversitas J. Biol. Divers. **25**, (2024)
20. M.M. Manee, B.M. Al-Shomrani, M.B. Al-Fageeh, Genome-wide characterization of simple sequence repeats in Palmae genomes, Genes Genom. **42**, 597–608 (2020)
21. P. Calabrese, R. Durrett, Dinucleotide Repeats in the Drosophila and Human Genomes Have Complex, Length-Dependent Mutation Processes, Mol. Biol. Evol. **20**, 715–725 (2003)
22. R.K. Kalia, M.K. Rai, S. Kalia, R. Singh, A.K. Dhawan, Microsatellite markers: an overview of the recent progress in plants, Euphytica. **177**, 309-334 (2011)

23. M. Zhao, G. Shu, Y. Hu, et al., Pattern and variation in simple sequence repeat (SSR) at different genomic regions and its implications to maize evolution and breeding, BMC Genomics. **24**, 136 (2023)
24. S. Liu, Y. An, F. Li, et al., Genome-wide identification of simple sequence repeats and development of polymorphic SSR markers for genetic studies in tea plant (*Camellia sinensis*), Mol. Breeding. **38**, 59 (2018)
25. M.Z. Iqbal, S. Jamil, R. Shahzad, K. Bilal, R. Qaisar, A. Nisar, M.K. Bhatti, DNA fingerprinting of crops and its applications in the field of plant breeding, J. Agric. Res. **59**, (2021)
26. K. Sunilkumar, P. Murugesan, R.K. Mathur, M.K. Rajesh, Genetic diversity in oil palm (*Elaeis guineensis* and *Elaeis oleifera*) germplasm as revealed by microsatellite (SSR) markers, Ind. J. Agric. Sci. **90**, 741-745 (2020)
27. M.R. Romadhon, Sobir, W.B. Suwarno, D.D. Matra, Development of Microsatellite Markers to Determine Genetic Diversity of Indonesian Betel Nut (Areca catechu L.), in Proceedings of the Nusantara Science and Technology Proceedings, Multi-Conference Proceeding Series D, January 5-13, (2023), 8-13
28. M.S. Rahayu, A. Setiawan, I. Maskromo, A. Purwito, S. Sudarsono, Genetic diversity analysis of Puan Kalianda kopyor coconuts (*Cocos nucifera*) from South Lampung, Indonesia based on SSR markers, Biodiversitas J. Biol. Divers. **23**, (2022)
29. D.Y. Rinawati, R. Reflinur, D. Dinarti, S. Sudarsono, Genetic diversity of sugar palm (*Arenga pinnata*) derived from nine regions in Indonesia based on SSR markers, Biodiversitas J. Biol. Divers. **22**, (2021)
30. J.A. Mantiquilla, M.S. Shiao, H.Y. Lu, K. Sridith, S.N.M. Sidique, W.K. Liyanage, Y.C. Chiang, Deep structured populations of geographically isolated nipa (*Nypa fruticans* Wurmb.) in the Indo-West Pacific revealed using microsatellite markers, Front. Plant Sci. **13**, 1038998 (2022)