BIODIVERSITAS
Volume 25, Number 3, March 2024
Pages: 1081-1088

ISSN: 1412-033X
E-ISSN: 2085-4722
DOI: 10.13057/biodiv/d250322

Profile microsatellite mining of whole genome sequencing and transcriptomic assembly in dwarf and tall areca nut (*Areca catechu*) in Indonesia

MUHAMMAD ROIYAN ROMADHON[1,♥], SOBIR[2,♥♥], WILLY BAYUARDI SUWARNO[2], DEDEN DERAJAT MATRA[2]

[1]Program of Plant Breeding and Biotechnology, Graduate School, Institut Pertanian Bogor. Jl. Raya Dramaga, IPB Dramaga Campus, Bogor 16680, West Java, Indonesia. Tel.: +62-251-8622961, ♥email: mroiyanripb1@gmail.com

[2]Department of Agronomy and Horticulture, Faculty of Agriculture, Institut Pertanian Bogor. Jl. Meranti, IPB Dramaga Campus, Bogor 16680, West Java, Indonesia. Tel./fax.: +62-251-8629353, ♥♥email:ridwaniisobir@gmail.com

Manuscript received: 6 October 2023. Revision accepted: 22 March 2024.

Abstract. *Romadhon MR, Sobir, Suwarno WB, Matra DD. 2024. Profile microsatellite mining of whole genome sequencing and transcriptomic assembly in dwarf and tall areca nut (*Areca catechu *L.) in Indonesia. Biodiversitas 25: 1081-1088.* Areca nut (*Areca catechu* L.) has high diversity in fruit shape and flowering time. Two superior varieties are commonly cultivated in Indonesia, namely Betara areca nut (tall) and Emas areca nut (dwarf). Molecular level diversity from genomic and transcriptome of areca nut data is available at NCBI, but not yet for Indonesian areca nut. The research compared the results of SSR with two method approaches to detect genetic diversity in the plants accurately. This study aimed to compare the SSR motifs of Betara areca nut and Emas areca nut varieties from whole genome sequencing and transcriptome assembly. The research was conducted at the Leuwikopo Seed Centre Laboratory of the IPB University, Bogor, Indonesia. The methods used are Whole Genome Sequencing (WGS) and transcriptome assembly. A total number of identified SSRs from WGS approach from Betara areca nut of 95 SSRs and Emas areca nut of 95 SSRs, while Betara areca nut SSRs from transcriptome assembly of 466 SSRs and Emas areca nut of 357 SSRs. The percentage of contigs and transcripts from Betara areca nut containing SSR was 7.43% and 4.22%, respectively, while Emas areca nut was 8.06% and 2.04%, respectively. AT is the dominant SSR motif in WGS, while the GA motif dominates the transcriptome assembly results.

Keywords: *Areca catechu,* coding sequence, single sequence repeats motif, transcriptome assembly, whole genome sequencing

INTRODUCTION

Areca nut (*Areca catechu* L.) belongs to the family Arecaceae (Palmae). Native range of this species is unknown, however it is a cultigen that exists only where humans grow it. Now, it is cultivated from East Africa and the Arabian Peninsula across tropical Asia and Indonesia to the central Pacific and New Guinea (Staples and Bevacqua 2006). Indonesia has areca germplasm scattered in regions, such as Gorontalo, Manado, Papua, North Kalimantan, West Kalimantan, and Jambi. Markers for studying genetic diversity use morphology, cytology, biochemistry, and molecular (DNA) (Nadeem et al. 2018).

Morphological markers based on plant phenotype have the advantage of being faster and cheaper, but are not consistent at plant development stages. Cytological markers to detect differences in chromosome pattern, number, size, and position. However, it is low polymorphisms, number of chromosomes, and low sensitivity (Kwiatek et al. 2019). The discovery of molecular markers interests researchers because they are stable in all plant phases in all tissues, with no environmental influences and epistatic effects. Molecular markers can assess diversity down to the DNA level and are stable in all phases (Salgotra and Stewart Jr 2019).

Simple Sequence Repeat (SSR) markers can be sequenced according to genetic level or Express Sequence Tag (EST). SSR markers help study genetics, assess genetic diversity, and utilize germplasm (Li et al. 2022). The markers are made by genomic and transcriptome data. The development of these markers and their polymorphic information represents a significant increase in available genomic resources. It will facilitate genetic and breeding applications, further accelerating the development of new cultivars (Wang et al. 2014). The advantages of SSR markers are low assay and equipment costs, high throughput rates, and ease of use (Ahmad et al. 2018). The development of useful SSR markers provides knowledge in research fields, such as genetic capture, comparative genome capture, and genomic-wide association analysis (Zhong et al. 2021). SSR is used as a clue to reveal the plant genome evolutions (Qin et al. 2015) and has been used for a variety of palm plants such as coconut (*Cocos nucifera*) (Rasam et al. 2016), oil palm (*Elaeis guineensis*) (Budiman et al. 2019), and date palm (*Phoenix dactylifera*) (Ahmed et al. 2021).

Moreover, the technological advances led to obtaining SSR using Next-Generation Sequencing (NGS). This approach can isolate SSRs from whole genome and transcriptome data starting with the denovo assembly stage and the final stage with SSR isolation so that repeat motifs can be obtained. The presence of NGS allows the in-silico approach

to become a cheap and fast plant genome sequencing innovation in developing SSR markers. NGS has high-throughput sequencing, resulting in large amounts of DNA or RNA sequencing (Goodwin et al. 2016). NGS has several steps for sequencing, for example, DNA sequencing involves DNA fragmentation, library preparation, massively parallel sequencing, bioinformatics, variant annotation, and interpretation (Qin 2019). NGS can be used for microsatellite isolation by Whole Genome Sequencing (WGS) and transcriptome assembly. SSR isolation using WGS has been carried out on cultivated spinach (*Spinacia oleracea*) (Bhattarai et al. 2021) and grape (*Vitis vinifera*) (Pei et al. 2023). SSR isolation for transcriptome assembly has been carried out in crispa (*Luctuca sativa*) (Zhang et al. 2021), oil palm (Zhou et al. 2020), red ginger (*Alpinia purpurata*) (Vidya et al. 2021), and golden saxifrage (*Chrysosplenium alternifolium*) (Xiang et al. 2023).

Related studies comparing SSR results on native Indonesian areca nut plants with whole genome sequencing and transcriptome assembly methods have never been carried out. This research can later be used to design SSR primers and detect the genetic diversity of areca nut. This study aimed to compare the SSR results of Betara areca nut and Emas areca nut varieties from WGS and transcriptome assembly.

MATERIALS AND METHODS

Plant materials

Samples for Whole Genome Sequencing (WGS) and transcriptome assembly were areca nut superior varieties in Indonesia, namely Emas areca nut and Betara areca nut. Each variety consists of two samples. Betara areca nut variety was obtained from Jambi, Indonesia. It was released by the Decree Minister of Agriculture No. 199/Kpts/SR. 120/1/2013 in 2013 as a superior variety. Description of this variety is as follows: age when flowering begins around 4-5 years, harvest age around 4-7 years, stem height around 10.28 m, number of fruit bunches/year is 5 fruit bunches per year, number of fruit per bunches 131.35, dry seed weight /grain of 8.68 g, potential dry kernels/tree/year of 5.70 kg, potential dry kernels/ha of 7.81 tons, tannin content of 9.79%. Emas areca nut variety was collected from Kotamobagu, North Sulawesi, Indonesia. It was released by the Decree of the Minister of Agriculture Number 39/KPTS/KB.020/2/2019 in 2019 as a superior variety. Description of this variety is as follows: the number of bunches per year of 5 fruit bunches per year, the number of fruit/bunches/trees of 75, and the weight of the fruit per seed is 65 g. Dry seed production per tree is 2.35 kg, and dry seed production/ha/year is 3.2 tons. The differences between the Emas areca nut and Betara areca nut include that Emas areca nut has shorter stems than Betara areca nut, the flowering age is 2-3 years and the harvest age is 3 years (Betara areca nut is 4-5 years).

Leaf samples with no pest or disease attack on the leaves were used for WGS analysis, while for transcriptome assembly used flower buds. All plant materials were obtained from the Kayuwatu Research Garden in Manado, North Sulawesi, Indonesia. DNA and RNA extraction activities were conducted at the Seed Center Laboratory of IPB University, Bogor, Indonesia from October 2022 to February 2023.

Procedures

Whole genome sequencing

DNA samples of Emas areca nut and Betara areca nut leaves were isolated using the CTAB protocol described by Doyle and Doyle (1990). The fresh leaf samples were taken approximately 0.4 g of leaf samples and were then crushed using 1.4 mL of lysis buffer, 0.01 g of PVP, and 4 µL of mercaptoethanol. The scour results were then put into a 2 mL tube and incubated in a water bath at 65°C for 60 minutes, inverted every 15 minutes. The sample was centrifuged using an Eppendorf Centrifuge 5416 at 12,000 rpm for 10 minutes. After that, the supernatant was transferred to a 1.5 mL tube, then add chloroform: isoamyl alcohol (24:1) as much as 1x the supernatant volume, and then inverted until mixed. The sample was then centrifuged at 12,000 rpm for 10 minutes. The supernatant was then taken out, and a single volume of chloroform was added once more. The DNA was cleaned by adding isoamyl alcohol and then centrifuged for 10 minutes at 12,000 rpm. The supernatant was taken and put into a 1.5 ml tube, then added isopropanol 0.8 x supernatant volume and NaCl 0.1 x supernatant volume. The samples were inverted and incubated in the freezer at -20°C until 24 hours. The sample was then centrifuged at 12,000 rpm for 10 minutes. The supernatant was then removed, and a DNA pellet was obtained. The pellet was washed using 70% ethanol, as much as 300 µL and inverted. The sample was then centrifuged at 12,000 rpm for 10 minutes, after which the supernatant was discarded, and the pellet was dried until no ethanol smell was emitted. The pellet was then dissolved using TE buffer 1x as much as 100 µL and stored in the freezer at -20°C until the DNA was used.

DNA quantification and quality tests were performed using a spectrophotometer with 2 µL of sample DNA. Results of DNA purity are seen from $\lambda 260/\lambda 280$. The steps of whole genome sequencing are library construction following standard protocols Oxford Nanopore Technologies (ONT) issued for DNA using the Q20+ Ligation Sequencing Kit (SQK-LSK114). The sequencing process used Flow Cell type R10.4.1 (FLO-MIN114D) on MinION Mk1B using MinION Release 21.11.7 software. Flye is a de novo assembler for single molecule sequencing readout (Galaxy Version 2.9.1+galaxy0).

Transcriptome assembly

RNA isolation with the RNeasy PowerPlant Kit (Qiagen). RNA quality and quantity were measured using the NanoPhotometer® NP80 (ImLen) and Qubit™ Fluorometer (Invitrogen). Library construction followed the standard protocol issued by the PCR-cDNA Barcoding Kit (SQK-PCB109). The sequencing process used Flow Cell type R9.4.1 (FLOMIN106D) on MinION Mk1B using MinION Release 21.11.7 software. Basecalling was processed with raw data using Guppy v6.0.1 software.

Data analysis

Denovo assembly results from WGS and Transcriptome assembly were performed with SSR isolation using the Misa Software program. The fragment containing SSR was isolated with Misa Software. Retrieval of SSR motifs refers to the method (Matra et al. 2021). Graphs and diagrams of SSR motifs produced by both methods were presented using Microsoft Excel.

RESULTS AND DISCUSSION

Microsatellites WGS and transcriptome assembly approaches

The search results for Betara areca nut and Emas areca nut microsatellites using the WGS approach showed the same results, with 95 identified SSRs. Total number of sequences examined of WGS from Betara areca nut was 888 sequences and 8,078 sequences from transcriptome assembly, whereas the number of sequences examined of WGS from Emas areca nut was 831 sequences and 15,040 sequences. The number of SSRs identified by the transcriptome assembly approach for Betara areca nut was 466 SSR, and Emas areca nut was 357 SSR. The number of sequences containing SSR with the WGS approach of Betara areca nut was 66 sequences, while for Emas areca nut was 67 sequences in contrast to the transcriptome assembly approach on Betara areca nut as many as 341 and Emas areca nut as many as 307. Number of sequences containing more than 1 SSR of Betara areca nut of WGS was 17 sequences and 62 sequences of transcriptome assembly, whereas number of sequences containing more than 1 SSR of WGS of Emas areca nut was 14 sequences and 28 sequences of transcriptome assembly. Number of SSRs present in the compound formation of WGS from Betara areca nut was 14 SSRs and 112 SSRs of transcriptome assembly, whereas Emas areca nut was 21 SSRs of WGS approach and 36 SSRs of transcriptome assembly (Table 1).

Distribution of SSR motifs

Betara areca nut SSR motifs from WGS results include 49.47% (dinucleotide), 45.26% (trinucleotide), and 5.26% (tetranucleotide). SSR motifs from Emas areca nut WGS results include 48.42% (dinucleotides), 46.32% (trinucleotides), and 5.26% (tetranucleotides). In contrast to the results of the SSR motif approach with transcriptome assembly, SSR motifs are higher than WGS. Betara areca nut SSR motifs include 83.69% (dinucleotide), 13.73% (trinucleotide), 2.15% (tetranucleotide), 0.21% (pentanucleotide), and 0.21% (hexanucleotide). Meanwhile, the SSR motifs in Emas areca nut include 74.51% (dinucleotide), 21.01% (trinucleotide), 3.92% (tetranucleotide), 0.28% (pentanucleotide), and 0.28% (hexanucleotide) (Figure 1).

SSR percentage of WGS and transcriptome assembly

A contig is a continuous sequence without gaps and is a former genome. The RNA transcript is the RNA strand that results from the transcribed gene. The SSR percentage was generated from the number of SSR-containing sequences and the total number of sequences examined. The number of SSR-containing sequences contains the number of SSRs in the sequence. The percentage of SSR from Betara areca nut and Emas areca nut showed more SSR resulting from the WGS approach of 7.43% and 8.06% respectively than the transcriptome assembly of 4.22% and 2.04% respectively (Figure 2).

Frequency of SSR motifs of WGS and transcriptome assembly

The frequency of SSR motifs from WGS in Betara areca nut was AT (14), TA (13), CT (5), GT (5), and TG (5) motifs (Figure 3.A). The frequency of SSR motifs from WGS results in Emas areca nut was AT (23), TA (12), TCT (5), TC (4), and AAG (4) (Figure 3.B). Different result in transcriptome assembly in Betara areca nut was GA (89), AG (64), AC (47), TG (45), and TC (31) (Figure 4.A). The SSR frequency resulting from transcriptome assembly in Emas areca nut was GA (50), AT (41), TC (32), TA (27), and CA (20) motifs (Figure 4.B). The result showed that the dominant SSR motif from genome data was AT, while the dominant SSR motif from transcriptome data was AG.

Correlation of SSR motif frequency from WGS and transcriptomics assembly

Exploring the relationship between the factors associated with the nucleotide yield of different WGS and transcriptomics approaches, we performed a correlation analysis for the factors. Transcriptome of Emas areca nut (TEA) is correlated with Transcriptome of Betara areca nut (TBA) of 0.92 and WGS of Emas areca nut (WEA) is correlated with WGS of Betara areca nut (WBA) of 0.82. TEA increases if TBA increases; if WEA increases, then WBA increases. These results indicate that each method for the two areca nut varieties is suitable because the two varieties are both positively correlated (Table 2).

Table 1. Microsatellite search results of Betara areca nut and Emas areca nut

Results of microsatellite search	Betara areca nut		Emas areca nut	
	Whole genome sequencing	Transcriptome assembly	Whole genome sequencing	Transcriptome assembly
Total number of sequences examined	888	8,078	831	15,040
Total size of examined sequences (bp)	4,264,762	4,894,902	3,825,274	4,449,267
Total number of identified SSRs	95	466	95	357
Number of SSR-containing sequences	66	341	67	307
Number of sequences containing more than 1 SSR	17	62	14	28
Number of SSRs present in compound formation	13	112	21	46

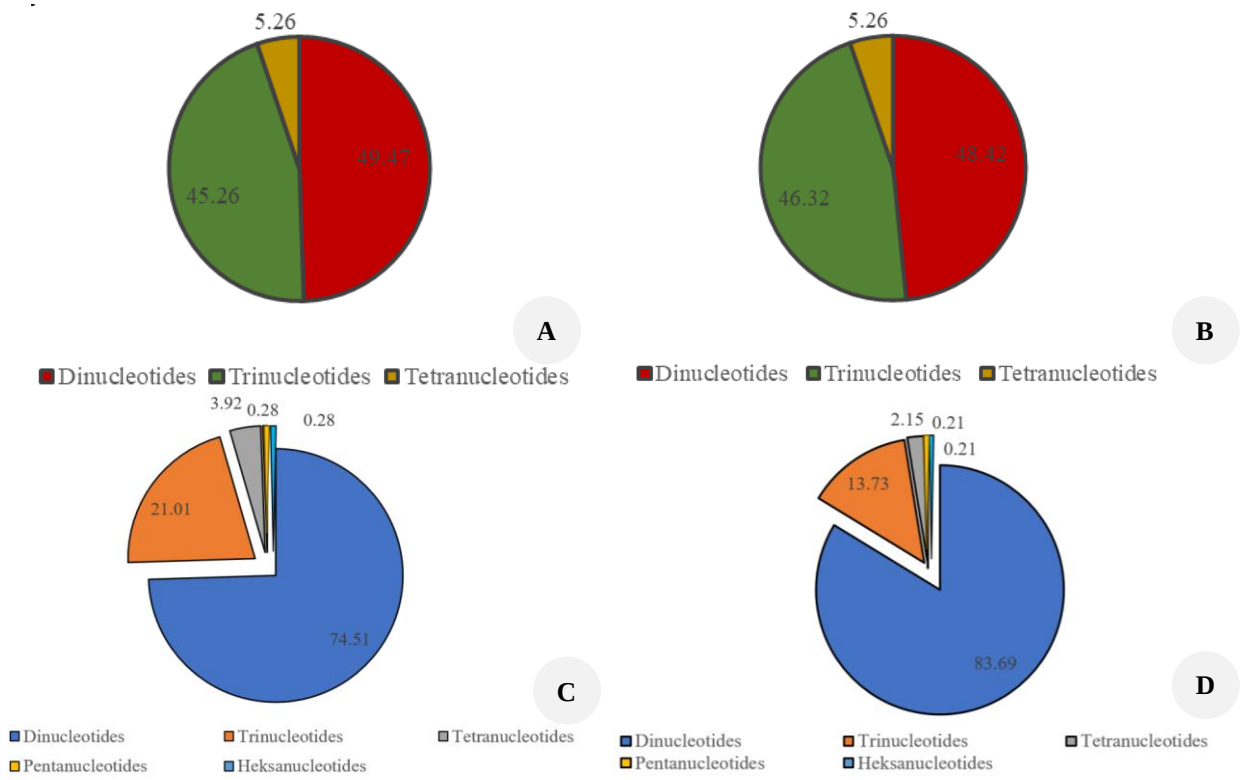

Figure 1. Distribution of SSR motifs from WGS: A. Betara areca nut, B. Emas areca nut, and from transcriptome assembly: C. Betara areca nut, D. Emas areca nut

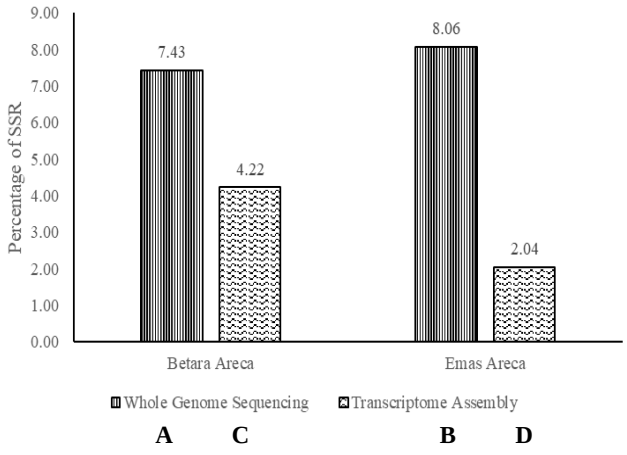

Figure 2. Percentage of SSR motifs from WGS: A. Betara areca nut, B. Emas areca nut, and from transcriptome assembly: C. Betara areca nut, D. Emas areca nut

Discussion

The genome sequences of plants from the Palmae family, such as *E. guineensis* and *P. dactylifera*, have been released; However, for areca nut plants, there are none, and the microsatellite results from WGS and transcriptome assembly have not yet been compared. There are 814,383 microsatellites from the whole genome of *E. guineensis* and 371,629 microsatellites from *P. dactylifera* (Xiao et al. 2016). There are 465 microsatellites of *E. guineensis* from transcriptome data (Tranbarger et al. 2012) and 5,981 microsatellites from *P. dactylifera* (Zhao et al. 2012). However, for our research, the results obtained for the two types of areca nut were smaller than previous studies because it used long reads sequencing techniques compared to previous studies, namely short reads sequencing. This causes the results from contigs (genome data) and transcripts (transcriptome data) to have longer read lengths per contig or transcript than short reads sequencing. In this research, the SSR results were possibly higher in the transcriptome assembly approach because introns previously separated the exons on CDS (coding sequence). However, the introns are sliced after transcription to connect the previously separated exons to a new sequence (transcript). In addition, the presence of isoforms increases the presence of SSR. RNA sequencing using a transcriptomics approach is the determination of bases in the form of active gene transcripts after transcription (Table 1). The number of SSRs present in compound formation, indicates that the SSR obtained is in the form of a combination of letters and numbers so it cannot be used as a diversity marker. Liu et al. (2021), Simple Sequence Repeat (SSR) is a DNA sequence repetition technique used in genetic research. The results of SSR isolation from SSRs made by transcriptomic SSRs are in the gene transcription region, while genomic SSRs are in the non-coding region of the genome. Yue et al. (2014), Expressed Sequence Tag SSRs (EST-SSRs), which reside in the transcribed regions of genes, are more evolutionarily conserved and more capable of transfer to related species than genomic SSRs. Compared with non-coding regions, genic regions have lower polymorphism due to functional limitations.

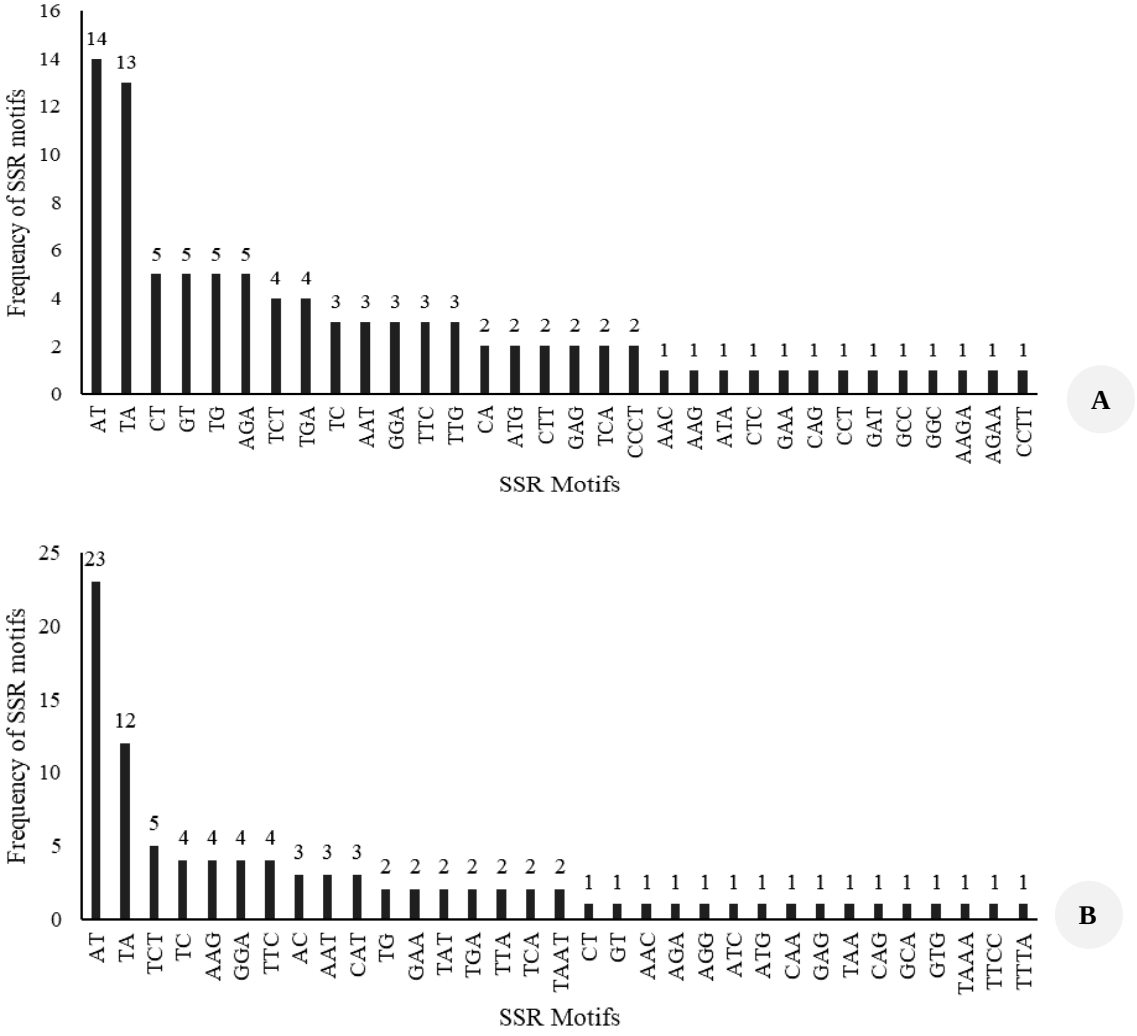

Figure 3. Frequency of SSR motifs identified by WGS of: A. Betara areca nut and B. Emas areca nut

Table 2. Correlation coefficient and p.value between SSR results from WGS and transcriptome assembly on Betara areca nut and Emas areca nut

	TEA	**TBA**	**WEA**	**WBA**
TEA	1.00			
Sig (2- tailed)	1.00			
TBA	0.92**	1.00		
Sig (2- tailed)	0.00	1.00		
WEA	0.36	0.10	1.00	
Sig (2- tailed)	039	0.42	1.00	
WBA	0.36	0.14	0.82**	1.00
Sig (2- tailed)	0.26	0.24	0.00	1.00

Note: TEA: Transcriptome of Emas Areca nut, TBA: Transcriptome of Betara Areca nut, WEA: WGS of Emas Areca nut, WBA: WGS of Betara Areca nut

The most abundant microsatellites in Betara areca nut and Emas areca nut from two approaches are dinucleotides (Figure 1). The dinucleotide SSR motif is the dominant motif in *Elaeis guineensis* at 76.6% and *Phoenix dactylifera* at 72.2%, using the WGS approach (Xiao et al. 2016) while the dominant motif from the transcriptome assembly of *Elaeis guineensis* is mononucleotide at 30.06 %, dinucleotides were 19.41%, tetranucleotides were 1.26%, pentanucleotides were 0.36% and hexanucleotides were 0.20% (Xiao et al. 2014). SSR microsatellite motifs include mononucleotides, dinucleotides, trinucleotides, tetranucleotides, pentanucleotides, and hexanucleotides. The mononucleotide SSR motif did not provide any significant meaning. The results of the whole genome sequencing approach only have dinucleotide, trinucleotide, and tetranucleotide motifs (Figures 1A and 1B). The results of the transcriptome assembly approach obtained pentanucleotide, hexanucleotide, dinucleotide, trinucleotide, and tetranucleotide SSR motifs. Zhao et al. (2023) reported SSR motifs and SSR variations in monocot plants have the pattern di->tri->tetra->penta->hexanucleotide P-SSRs, and this is the same as the pattern in areca palm plants with the smallest distribution of SSR motifs, namely pentanucleotides and hexanucleotides ranged from 0.21-0.28% (Figures 1C and 1D). The most common distribution of the two approaches is the dinucleotide SSR motif. The dominant SSR motifs in other palmae families are dinucleotides in coconut (*Cocos nucifera*) (Hatta et al. 2022)

and oil palm (*Elaeis guineensis*) (Xiao et al. 2014; Zhou et al. 2020). SSR dinucleotide motifs are dominant in other plants such as Oil-Camellia (*Camellia oleifera*) (Tian et al. 2022) and Himalayan Canary (*Juglans regia*) (Itoo et al. 2023). The abundance of dinucleotides increases the chance of amplification, because there are many dinucleotides compared to the others. The abundance of dinucleotides is a sign of each genomic DNA, which can distinguish between sequences from different organisms (Karlin and Burge 1995). SSR arrangements can be classified by motif as: (i) perfect when composed entirely of repetitions of a single motif; (ii) imperfect if a base pair does not belong to a motif that occurs between repetitions; (iii) interrupted if a sequence of several base pairs is inserted into the motif; or (iv) combined if formed by several adjacent and repeating motifs (Abdurakhmonov and Abdukarimov 2008).

SSR percentages of Betara areca nut and Emas areca nuts revealed that the WGS method produced higher SSR (7.43% and 8.06%) than the transcriptome assembly method (4.22% and 2.04%) (Figure 2). The SSR percentage of *Elaeis guineensis* using the WGS approach was 20.96% (Xiao et al. 2016), while the SSR percentage of the transcriptome assembly approach was 9.78% (Xiao et al. 2014). In this case, the genome data has a ratio between the number of SSR-containing sequences and the total number of sequences examined that is greater than the transcriptome data.

The AT motif is the dominant motif in the WGS approach and transcriptome assembly of *Elaeis guineensis* (Xiao et al. 2014). Qu and Liu (2013) found that about 80% of GC-containing trinucleotides are in exons, while AT-rich trinucleotides are almost evenly distributed throughout all components of the genome (untranslated regions, intergenic spaces, and introns). Tetranucleotides are primarily found in the non-coding regions of the rice genome, especially the intergenic regions. In general, SSRs have the highest mutation rates related to gene expression, and mutations in these regions contain the lowest SSRs within the gene region. SSRs with trinucleotide and hexanucleotide motifs in the coding area are often found in mutations that change the reading frame (Zhang et al. 2004; Xu et al. 2013).

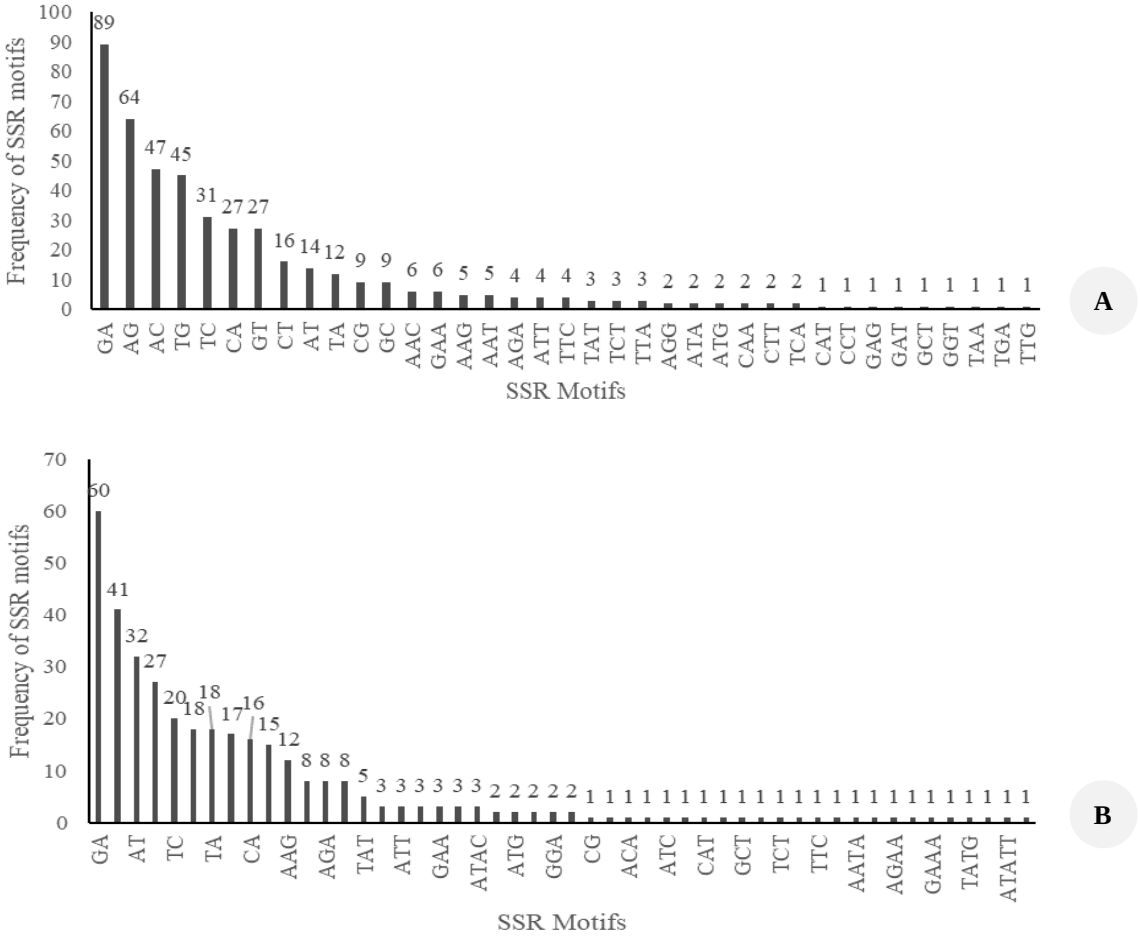

Figure 4. Frequency of SSR motifs identified by transcriptome assembly of: A. Betara areca nut and B. Emas areca nut

The promoter of the TATA box gene is occupied by a regulatory molecule that is different from the molecule that binds to the promoter of the TATA-free gene (Savinkova et al. 2023). Qin et al. (2015), Evolution in plant genomes, the repetition of nucleotide motifs is highest in AT or TA nucleotide combinations. AT is widely found in WGS because it contains UTR (Untranslated Region), namely promoter, terminator, and gene parts (Xiao et al. 2016). The gaps in the gene may contain many AT repeats. In addition, the promoters are regions that contain AT-rich, because they are transcriptional region (TATA boxes) attachment sites. Promoter gene, transcriptional region, and terminator; if this gene is regulated, the gene will experience on and off. In genes that experience on and off, genes are attached to the promoter. The protein that deactivates the gene is a transcriptional factor and attaches to the promoter (cis-acting element). The cis-acting element is the binding site for trans-acting factors, such as transcription factors or long noncoding RNAs. The cis-acting elements play an important role in the regulation of gene expression, development, and evolutionary processes (Kaur et al. 2017). Polymorphisms found in SSR are variations in template number repeats induced by DNA replication or recombination errors due to polymerase strand slippage due to defects in DNA replication due to incompatibility between a template and new strands. Recombination events, such as unequal crossing and gene transfer, can expand SSR sequences. The purer and longer the repeats, the higher the risk of mutation, whereas shorter repeats with low purity suffer lower mutation frequency. Mutations generate new alleles at the SSR locus through DNA mismatch repair mechanisms (Zhu and Yu 2009).

The correlation coefficient is a value used to measure the degree of closeness of the relationship between two variables (Altman and Krzywinski 2015; Chatterjee 2020; Janse et al. 2021). The correlation between the two variables is quantified with numbers between −1 and +1. Zero means no correlation, whereas 1 means complete or perfect correlation. A negative R-value indicates an inverse relationship. The strength of the correlation increases from 0 to +1 (direct relationship) or 0 to -1 (inversely strong relationship) (Akoglu 2018). The correlation results from the two methods are above 0.5, which means the correlation is positive and strong. The relationship between TEA and TBA and WEA with WBA is strong (Table 2). So, if there is an increase in SSR results from TEA, it will cause an increase in SSR results in TBA. A positive and strong correlation means that each method is stable in producing SSR for both areca nut varieties.

In conclusion, A total of 95 SSRs from Betara areca nut and 95 SSRs from Emas areca nut were found using the WGS approach, whereas 466 SSRs from Betara areca nut and 357 SSRs from Emas areca nut were found using transcriptome assembly. The percentage of contigs and transcripts from Betara areca nut containing SSR was 7.43% and 4.22%, respectively, while Emas areca nut was 8.06% and 2.04% respectively. AT is the dominant SSR motif in WGS, while GA motif dominates the transcriptome assembly results. The possibility of SSR results is more significant in the transcriptomic approach, because in the gene sequence previously, the exons on CDS (coding sequence) were separated by introns, but after transcription, a slicing process occurs on the introns so that previously separate exons are connected to become a new sequence (transcript) or there is an isoform that increases the presence of SSR The most SSR motif for WGS is AT, while for transcriptomics is GA. The sample used for transcriptomics is flower buds, so the organs taken from the buds will be different from other organs that will be sampled. The existence of these introns affects the creation of an SSR database using transcriptome data.

ACKNOWLEDGEMENTS

This research was funded by the Ministry of Research, Technology and Higher Education, Republic of Indonesia, through the Doctoral Dissertation Research second years funding on the 2023 scheme grant given to Prof. Dr. Ir Sobir, M.Si, and the team with contract number 001/E5/PG.02.00.PL/2023.

REFERENCES

Abdurakhmonov IY, Abdukarimov A. 2008. Application of association mapping to understanding the genetic diversity of plant germplasm resources. Intl J Plant Genomics 2008: 574927. DOI: 10.1155/2008/574927.

Ahmad A, Wang J-D, Pan Y-B, Sharif R, Gao S-J. 2018. Development and use of Simple Sequence Repeats (SSRs) markers for sugarcane breeding and genetic studies. Agronomy 8 (11): 260. DOI: 10.3390/agronomy8110260.

Ahmed W, Feyissa T, Tesfaye K, Farrakh S. 2021. Genetic diversity and population structure of date palms (*Phoenix dactylifera* L.) in Ethiopia using microsatellite markers. J Genet Eng Biotechnol 19 (1): 64. DOI: 10.1186/s43141-021-00168-5.

Akoglu H. 2018. User's guide to correlation coefficients. Turk J Emerg Med 18 (3): 91-93. DOI: 10.1016/j.tjem.2018.08.001.

Altman N, Krzywinski M. 2015. Association, correlation and causation. Nat Methods 12 (10): 899-900. DOI: 10.1038/nmeth.3587.

Bhattarai G, Shi A, Kandel DR, Solís-Gracia N, da Silva JA, Avila CA. 2021. Genome-wide Simple Sequence Repeats (SSR) markers discovered from whole-genome sequence comparisons of multiple spinach accessions. Sci Rep 11 (1): 9999. DOI: 10.1038/s41598-021-89473-0.

Budiman LF, Apriyanto A, Pancoro A, Sudarsono. 2019. Illegitimacy testing of *Elaeis guineensis* population based on simple sequence repeat markers. Agrivita J Agric Sci 41 (3): 504-512. DOI: 10.17503/agrivita.v41i3.1969.

Chatterjee S. 2020. A new coefficient of correlation. J Am Stat Assoc 116 (536): 2009-2022. DOI: 10.1080/01621459.2020.1758115.

Doyle JJ, Doyle JL. 1990. Isolation of plant DNA from fresh tissue. Focus 12 (1): 13-15.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: Ten years of next-generation sequencing technologies. Nat Rev Genet 17 (6): 333-351. DOI: 10.1038/nrg.2016.49.

Hatta ANNL, Sukma D, Maskromo I, Sudarsono. 2022. Mining and validating novel SSR markers based on coconut (*Cocos nucifera* L.) whole genome and their use for phylogenetic analysis. Biodiversitas 23 (10): 5122-5131. DOI: 10.13057/biodiv/d231019.

Itoo H, Shah RA, Qurat S, Jeelani A, Khursheed S, Bhat ZA, Mir MA, Rather GH, Zargar SM, Shah MD, Padder BA. 2023. Genome-wide characterization and development of SSR markers for genetic diversity analysis in northwestern Himalayas walnut (*Juglans regia* L.). 3 Biotech 13 (5): 136. DOI: 10.1007/s13205-023-03563-6.

Janse RJ, Hoekstra T, Jager KT, Zoccali C, Tripepi G, Dekker FW, van Diepen M. 2021. Conducting correlation analysis: Important limitations and pitfalls. Clin Kidney J 14 (11): 2332-2337. DOI: 10.1093/ckj/sfab085.

Karlin S, Burge C. 1995. Dinucleotide relative abundance extremes: A genomic signature. Trends Genet 11 (7): 283-290. DOI: 10.1016/s0168-9525(00)89076-9.

Kaur A, Pati PK, Pati AM, Nagpal AK. 2017. In-silico analysis of cis-acting regulatory elements of pathogenesis-related proteins of *Arabidopsis thaliana* and *Oryza sativa*. PLoS One 12 (9): e0184523. DOI: 10.1371/journal. pone.0184523.

Kwiatek MT, Kurasiak-Popowska D, Mikołajczyk S, Niemann J, Tomkowiak A, Weigt D, Nawracała J. 2019. Cytological markers used for identification and transfer of *Aegilops* spp. chromatin carrying valuable genes into cultivated forms of *Triticum*. Comp Cytogenet 13 (1): 41-59. DOI: 10.3897/compcytogen.v13i1.30673.

Li X, Qiao L, Chen B, Zheng Y, Zhi C, Zhang S, Pan Y, Cheng Z. 2022. SSR markers development and their application in genetic diversity evaluation of garlic (*Allium sativum* L.) germplasm. Plant Divers 44 (5): 481-491. DOI: 10.1016/j.pld.2021.08.001.

Liu H, Zhang Y, Wang Z, Su Y, Wang T. 2021. Development and application of EST-SSR markers in *Cephalotaxus oliveri* from transcriptome sequences. Front Genet 12: 759557. DOI: 10.3389/fgene.2021.759557.

Matra DD, Fathoni MAN, Majiidu M, Wicaksono H, Sriyono A, Gunawan G, Susanti H, Sari R, Fitmawati F, Siregar IZ, Widodo WD, Poerwanto R. 2021. The genetic variation and relationship among the natural hybrids of *Mangifera casturi* Kosterm. Sci Rep 11 (1): 19766. DOI: 10.1038/s41598-021-99381-y.

Nadeem MA, Nawaz MA, Shahid MQ, Doğan Y, Comertpay G, Yıldız M, Hatipoğlu R, Ahmad F, Alsaleh A, Labhane N, Özkan H, Chung G, Baloch FS. 2018. DNA molecular markers in plant breeding: Current status and recent advancements in genomic selection and genome editing. Biotechnol Biotechnol Equip 32 (2): 261-285. DOI: 10.1080/13102818.2017.1400401.

Pei D, Song S, Kang J, Zhang C, Wang J, Dong T, Ge M, Pervaiz T, Zhang P, Fang J. 2023. Characterization of Simple Sequence Repeat (SSR) markers mined in whole grape genomes. Genes 14 (3): 663. DOI: 10.3390/ genes14030663.

Qin D. 2019. Next-generation sequencing and its clinical application. Cancer Biol Med 16: 4-10. DOI: 10.20892/j.issn.2095-3941.2018.0055.

Qin Z, Wang Y, Wang Q, Li A, Hou F, Zhang L. 2015. Evolution analysis of simple sequence repeats in plant genome. PLoS One 10 (12): e0144108. DOI: 10.1371/journal.pone.0144108.

Qu J, Liu J. 2013. A genome-wide analysis of simple sequence repeats in maize and the development of polymorphism markers from next-generation sequence data. BMC Res Notes 6: 403. DOI: 10.1186/1756-0500-6-403.

Rasam DV, Gokhale NB, Sawardekar SV, Patil DM. 2016. Molecular characterisation of coconut (*Cocos nucifera* L.) varieties using ISSR and SSR markers. J Hortic Sci Biotechnol 91 (4): 347-352. DOI: 10.1080/14620316.2016.1160544.

Salgotra RK, Stewart Jr NC. 2020. Functional markers for precision plant breeding. Intl J Mol Sci 21 (13): 4792. DOI: 10.3390/ijms21134792.

Savinkova LK, Sharypova EB, Kolchanov NA. 2023. On the role of tata boxes and tata-binding protein in *Arabidopsis thaliana*. Plants 12 (5): 1000. DOI: 10.3390/plants12051000.

Staples GW, Bevacqua RF. 2006. *Areca catechu* (betel nut palm). Species Profiles for Pacific Island Agroforestry. www.traditionaltree.org.

Tian Q, Huang B, Huang J, Wang B, Dong L, Yin X, Gong C, Wen Q. 2022. Microsatellite analysis and polymorphic marker development based on the full-length transcriptome of *Camellia chekiangoleosa*. Sci Rep 12 (1): 18906. DOI: 10.1038/s41598-022-23333-3.

Tranbarger TJ, Kluabmongkol W, Sangsrakru D, Morcillo F, Tregear JW, Tragoonrung S, Billotte N. 2012. SSR markers in transcripts of gene linked to post-transcriptional and transcriptional regulatory functions during vegetative and reproductive development of *Elaeis guineensis*. BMC Plant Biol 12: 1. DOI: 10.1186/1471-2229-12-1.

Vidya V, Prasath D, Snigdha M, Gobu R, Sona C, Maiti CS. 2021. Development of EST-SSR markers based on transcriptome and its validation in ginger (*Zingiber officinale* Rosc.). PLoS One 16 (10): e0259146. DOI: 10.1371/journal. pone.0259146.

Wang S, Liu Y, Ma L, Liu H, Tang Y, Wu L, Wang Z, Li Y, Wu R, Pang X. 2014. Isolation and characterization of microsatellite markers and analysis of genetic diversity in chinese jujube (*Ziziphus jujuba* Mill.). PLoS One 9 (6): e99842. DOI: 10.1371/journal.pone.0099842.

Xiang N, Lu B, Yuan T, Yang T, Guo J, Wu Z, Liu H, Liu X, Qin R. 2023. De novo transcriptome assembly and EST-SSR marker development and application in *Chrysosplenium macrophyllum*. Genes 14 (2): 279. DOI: 10.3390/genes14020279.

Xiao Y, Xia W, Ma J, Mason AS, Fan H, Shi P, Lei X, Ma Z, Peng M. 2016. Genome-wide identification and transferability of microsatellite markers between Palmae species. Front Plant Sci 7: 1578. DOI: 10.3389/fpls.2016.01578.

Xiao Y, Zhou L, Xia W, Mason AS, Yang Y, Ma Z, Peng M. 2014. Exploiting transcriptome data for the development and characterization of gene-based SSR markers related to cold tolerance in oil palm (*Elaeis guineensis*). BMC Plant Biol 14: 384. DOI: 10.1186/s12870-014-0384-2.

Xu J, Liu L, Xu Y, Chen C, Rong T, Ali F, Zhou S, Wu F, Liu Y, Wang J, Cao M, Lu Y. 2013. Development and characterization of simple sequence repeat markers providing genome-wide coverage and high resolution in maize. DNA Res 20: 497-509. DOI: 10.1093/dnares/dst026.

Yue X-Y, Liu G-Q, Zong Y, Teng Y-W, Cai D-Y. 2014. Development of genic SSR markers from transcriptome sequencing of pear buds. J Zhejiang Univ Sci B 15 (4): 303-312. DOI: 10.1631/jzus.b1300240.

Zhang C, Wu Z, Jiang X, Li W, Lu Y, Wang K. 2021. De novo transcriptomic analysis and identifcation of EST-SSR markers in *Stephanandra incisa*. Sci Rep 11: 1059. DOI: 10.1038/s41598-020-80329-7.

Zhang L, Yuan D, Yu S, Li Z, Cao Y, Miao Z, Qian H, Tang K. 2004. Preference of simple sequence repeats in coding and non-coding regions of *Arabidopsis thaliana*. Bioinformatics 20 (7): 1081-1086. DOI: 10.1093/bioinformatics/bth043.

Zhao M, Shu G, Hu Y, Cao G, Wang Y. 2023. Pattern and variation in Simple Sequence Repeat (SSR) at diferent genomic regions and its implications to maize evolution and breeding. BMC Genomics 24 (1): 136. DOI: 10.1186/s12864-023-09156-0.

Zhao Y, Williams R, Prakash CS, He G. 2012. Identification and characterization of gene-based SSR markers in date palm (*Phoenix dactylifera* L.). BMC Plant Biol 12: 237. DOI: 10.1186/1471-2229-12-237.

Zhong Y, Cheng Y, Ruan M, Ye Q, Wang R, Yao Z, Zhou G, Liu J, Yu J, Wan H. 2021. High-throughput ssr marker development and the analysis of genetic diversity in *Capsicum frutescens*. Horticulturae 7 (7): 187. DOI: 10.3390/ horticulturae7070187.

Zhou L, Yarra R, Zhao Z, Jin L, Cao H. 2020. Development of SSR markers based on transcriptome data and association mapping analysis for fruit shell thickness associated traits in oil palm (*Elaeis guineensis* Jacq.). 3 Biotech 10 (6): 280. DOI: 10.1007/s13205-020-02269-3.

Zhu C, Yu J. 2009. Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. Genetics 182 (3): 875-888. DOI: 10.1534/genetics.108.098863.