Journal of Oil Palm Research
DOI: https://doi.org/10.21894/jopr.2022.0008

OIL PALM SSR RESOURCE INTERFACE (OPSRI) – WEB-BASED BIOINFORMATIC ANALYSIS PIPELINE FOR SSR MINING

ROZANA ROSLI[1]*; MOHD AMIN AB HALIM [1]; TING NGOOT-CHIN[1]; NOORHARIZA MOHD ZAKI[1]; JAYANTHI NAGAPPAN[1]; RAJINDER SINGH[1]; LESLIE LOW ENG-TI[1] and ZETI-AZURA MOHAMED-HUSSEIN[2,3]

ABSTRACT

Co-dominant simple sequence repeat (SSRs) are popular DNA markers, widely applied in oil palm genetic studies, especially in genetic mapping, marker-trait association, diversity analysis and detecting illegitimacy. A repository having detailed information on SSRs and their inheritance profiles in specific breeding lines is lacking in Malaysia. Such a database enables prioritising polymorphic SSRs for screening, identification of which otherwise can be expensive and time consuming. As such, to facilitate and accelerate development of informative SSRs, oil palm SSR resource interface (OPSRI) was established and currently contains information on 1983 markers that were genotyped successfully across several oil palm breeding lines. OPSRI is a well-structured database that can expedite marker development and reduce redundancy in the on-going efforts at developing DNA markers for genetic analysis of oil palm. The system has been integrated with MISA, Primer3, ORF search script and BLAST to enable SSR identification. This study also characterised a large number of expressed sequence tags (EST)-SSRs using an in-silico *approach. The information within the OPSRI database can help accelerate research in oil palm as well as in crops such as coconut and date palm which have a high level of synteny and marker-transferability with oil palm.*

Keywords: *Elaeis guineensis*, ESTs, SSR database.

Received: 6 January 2021; **Accepted:** 9 December 2021; **Published online:** 28 January 2022.

INTRODUCTION

Projects related to the generation of expressed sequence tags (ESTs) for understanding the transcribed space in the genomes of many plant species have aided gene identification and the discovery of simple sequence repeat (SSR) markers. SSRs are short sequences of DNA that are repeated 1-6 times (Bhattarai *et al*., 2021), and their identification and utilisation have been reported for many plants (Al-Faifi *et al*., 2016; Song *et al*., 2016; Vieira *et al*., 2016). SSR markers have contributed immensely to the development and construction of genetic and physical maps. Among the wide repertoire of markers available for oil palm, SSRs have advantages as they are the most polymorphic co-dominant markers, high reproducibility, highly abundant in the genome and can be assayed using high resolution agarose gel electrophoresis systems that can be implemented in-house by many oil palm research stations (Gupta *et al*., 1999; Singh *et al*., 2007; Tautz and Renz, 1984; Zolkafli *et al*., 2021). Interestingly, the location of an SSR locus in the genome has an influence on its level of

[1] Malaysian Palm Oil Board,
6 Persiaran Institusi, Bandar Baru Bangi,
43000 Kajang, Selangor, Malaysia.

[2] Department of Applied Physics,
Faculty of Science and Technology
Universiti Kebangsaan Malaysia,
43600 UKM Bangi, Selangor, Malaysia.

[3] Centre for Bioinformatics Research,
Institute of Systems Biology (INBIOSIS)
Universiti Kebangsaan Malaysia,
43600 UKM Bangi, Selangor, Malaysia.

* *Corresponding author e-mail: lizana@mpob.gov.my*

polymorphism and specific function in a species (Lebedev *et al*., 2020), where for example SSRs in the 5'UTR are more polymorphic compared to those in the 3'UTR (Wan *et al*., 2020). SSRs found in the 5'UTR have an effect on gene transcription and regulation, whereas SSRs found in the 3'UTR are involved in gene silencing and transcription slippage (Li *et al*., 2004; Varshney *et al*., 2005). Furthermore, SSR markers located in non-coding introns (genomic SSRs), although not transcribed, are known to be involved in regulatory functions that influence plant development (Bagshaw, 2017; Lebedev *et al*., 2020; Tranbarger *et al*., 2012), thus, increasing their potential utility and informativeness. As such, determining the position of the SSR, whether it is in the 3'UTR, 5'UTR, coding or non-coding region is useful to further categorise the SSR. The information can be used to preselect and prioritise SSR markers for use in genetic analysis of the selected crop. Moreover, SSRs in genic regions are abundant and especially useful for tagging specific traits using the candidate gene approach.

In oil palm, SSRs are popular choice, and several studies have demonstrated the utilisation and efficiency of these markers. SSR markers have proven to be effective for use in studies such as genetic mapping, marker-trait association, as well as for analysing diversity of germplasm and advanced breeding lines (Bhagya *et al*., 2020; Sunilkumar *et al*., 2020; Ting *et al*., 2013). More recently, Sarimana *et al.* (2021) demonstrated the effectiveness of these markers as tools for DNA fingerprinting, while Zolkafli *et al.* (2021) identified a core set of SSR markers that can be routinely used for the same purpose in a wide variety of oil palm genetic backgrounds. Previous studies have also established cross transferability of SSR (including EST-SSR) markers across closely related plants species, which enables a better understanding of their evolutionary history. Zaki *et al.* (2012) demonstrated that the SSR markers from oil palm could amplify across species and genera of the family Arecaceae. Meanwhile Bazzo *et al.* (2018), also observed a high transferability rate (>70%) of macauba palm-derived EST-SSR markers across six palm species in the Arecaceae family, namely; *Acrocomia intumescens*, *Acrocomia totai*, *E. guineensis*, *Bactris gasipaes*, *Euterpe edulis* and *Sabal causiarum*. Similarly reported for other plant species such as *Linum* (Soto-Cerda *et al*., 2011), banana (Backiyarani *et al*., 2013), *Prunus* (Sorkheh *et al*., 2016) and chrysanthemum (Fan *et al*., 2019). Transferability across closely related species and within genera facilitates genetic research, especially in understanding the mutational processes that have taken place within these regions among closely related plant species (Zaki *et al*., 2012).

Although SSR markers have been widely utilised in oil palm research due to their ease of use at relatively low cost and availability from difference sources, a comprehensive database resource is still lacking. Two SSR databases are available in the public domain for oil palm: OpSatdb https://ssr.icar.gov.in/index.php (Babu *et al*., 2019) and TropGENE-DB https://tropgenedb.cirad.fr/tropgene/JSP/index.jsp (Hamelin *et al*., 2012), but these databases are limited to only listing the SSR markers. An expanded database that integrates experimental information related to the SSR markers and the potential utility of the SSR markers, will be more useful and desirable to researchers. The resource is required to help mine and characterise these markers more efficiently from the ever-growing repertoire of genomic resources that are becoming available. The huge amounts of data already available and new information being generated from the sequencing of additional oil palm breeding and germplasm lines have led to challenges in *de novo* mining of SSR and their utilisation. The existing EST collection and increasing availability of transcriptome and other related sequences, suggest that mining of SSRs from this enormous resource requires considerable technical skills, time and cost for successful execution. A number of software and scripts are available to assist in identifying SSR in sequences such as FullSSR (Metz *et al*., 2016), GMATA (Wang and Wang, 2016), Krait (Du *et al*., 2018) and recently reported Simple Sequence Repeat Molecular Marker Developer (SSRMMD) (Gou *et al*., 2020). Nevertheless, data mining of SSRs would be more practical and cheaper when retrieved from an organised database (Vieira *et al.* 2016), especially one with detailed information on the polymorphism of the SSR markers in selected breeding lines. The main limitation currently is that the SSR data and other relevant information, especially for oil palm, are mostly archived using spreadsheets, which are not robust and data retrieval can be difficult and messy. It is also difficult to integrate analytical tools to the spreadsheets. The availability of an automated database system, with easy-to-use analysis tools will enable researchers to more effectively retrieve relevant SSRs from ESTs and other collections of oil palm sequences for the development of SSR (including EST-SSR) markers. The availability of a web-based data management system will also make it more practical to manage information related to these markers such as motif type, repeat length, and position in the genome. Adding experimental information related to the specific SSRs, such as polymorphism rates and inheritance patterns in specific genetic backgrounds can expedite future research activities utilising these markers, especially in studies related to genetic mapping, QTL and diversity analysis.

MATERIALS AND METHODS

Data Source and Analysis

The dataset used in this study includes information on the sequence of the EG5.1 scaffold as well as amplification and segregation patterns of SSR markers in specific oil palm families. The information is stored in the database, where the user-friendly web interface facilitates data browsing. The data can also be downloaded easily through query interface. The experimental marker data-sets were generated from three mapping families of different genetic backgrounds namely, P2, T128 and OxG. P2 is a mapping family consisting of 87 *tenera* palms from Ulu Remis Deli *dura* (ENL48) and a Yangambi *pisifera* (ML161) (Ting *et al.*, 2013; 2014). T128 is a family consisting of 241 individual palms generated by self-pollination of a Nigerian *tenera* palm coded as 0.151/128, which is part of MPOB's germplasm collection (Singh *et al.*, 2013). The 0.151/128 was reported to have very high unsaturated oil with an iodine value (IV) of 63.4 (Kushairi *et al.*, 2011), which is higher than that observed in commercial *E. guineensis* palms. The OxG interspecific family consists of 108 F_1 hybrids generated by crossing the Colombian *E. oleifera* (UP1026) and palm 0.151/128 (Singh *et al.*, 2009; Ting *et al.*, 2014).

Development of the Oil Palm SSR Resource Interface (OPSRI)

OPSRI was developed using HTML5, PHP scripting language, CSS3 styling code, Javascript and Bootstrap framework packages. The MariaDB database is used to manage the SSR data. *Figure 1* shows an Entity-Relationship Diagram (ERD) representing the database module. The ERD clearly shows the two main independent entities, namely the SSR markers in the public (published) and private (unpublished) databases, with two normalised tables, namely the library and clone_mypalmviewer tables. Markers from public and private entities may have been derived from the sequencing of the same library information which is available in the library table. The information which is available in the library table describes library details such as breeding line, genotype, species and methylation status. All SSR markers were mapped to oil palm genome build EG5.1 and their genome locations are stored in the clone_mypalmviewer table. A collection of experimental results obtained from the testing of the oil palm SSR primers on specific populations and/or breeding lines is deposited in primer_exp_res table. The specific populations and/or breeding lines for which marker data is available currently are the P2, T128 and OG families, which have been described above. In addition, the information on the location of the SSR containing sequences on the published oil palm genome build (EG5.1) is available in clone_mypalmviewer and can be visualised in Oil Palm Genome Browser: MyPalmViewer, http://gbrowse.mpob.gov.my (Low *et al.*, 2020). The architecture of the system is important as it provides information on the dataflow and shows efficiency of data handling. The architecture is flexible and can accommodate integration of new systems or any restructuring required in the future. Apache servers are used to enable interaction between users and applications. This system is not restricted to any one operating system (OS) as it was tested on Windows, LINUX and Mac OS. The analysis pipeline in OPSRI was developed using open-source bioinformatics tools such as NCBI BLAST (Altschul *et al.*, 1997), MISA scripts (Thiel *et al.*, 2003) and Primer3 programmes (Rozen and Skaletsky, 2000) and ORF search scripts (*Figure 2*).

In silico Mining of SSRs

Oil palm EST sequences, downloaded from NCBI GenBank database were assembled using the CAP3 assembly programme (Huang and Madan, 1999), resulting in contigs with consensus sequences and singletons, which help to avoid redundancy. The batch FASTA file of consensus and singleton sequences was used to mine for SSRs using MISA (MIcroSAtellite) (Thiel *et al.*, 2003). Criteria used for the SSR mining were minimum of 10 repeats for mononucleotide motifs, 6 repeats for dinucleotide, 5 repeats for trinucleotide, 4 for tetranucleotide and 5 for pentanucleotide repeat motifs. The maximum number of interrupting bases between two SSRs in a compound microsatellite was less than 100 bp. Additional bash scripts used the MISA output file as an input file to design primer pairs using Primer3 (Rozen and Skaletsky, 2000). Parameters set for primer design were as follows: length range from 18-24 bp, melting temperature 57°C-62°C and expected fragment size of 100-300 bp.

Classification of SSR Distribution

The SSR sequences were further classified into three categories, based on their genomic positions, 5'UTR, 3'UTR and coding region. The classifications were done by identifying the full-length open reading frames (ORF) of the unique gene datasets. SSRs located before the start codon were classified as 5' UTR whereas, SSRs found after the stop codon were considered to be in the 3' UTR. Finally, repeat motifs located between the start and stop codons were considered as SSRs located within the coding sequence (CDS) region. Blast analysis was also conducted on the SSR containing sequences, against the RefSeq protein database (O'Leary *et al.*, 2016) using NCBI BLAST (Altschul *et al.*, 1997). All positive hits with an e-value lower than 1×10^{-6} were included in the analysis.

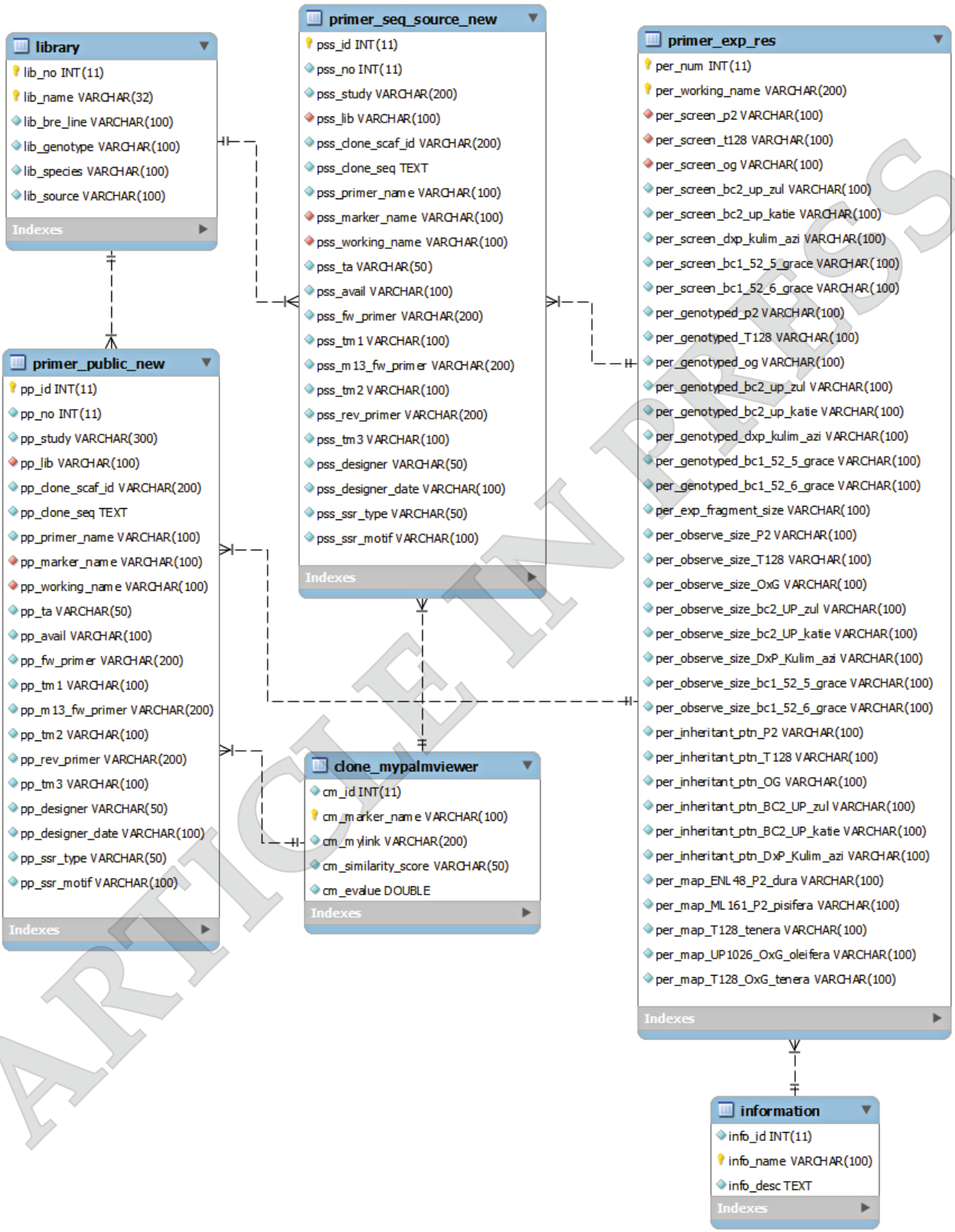

Figure 1. ERD of OPSRI database.

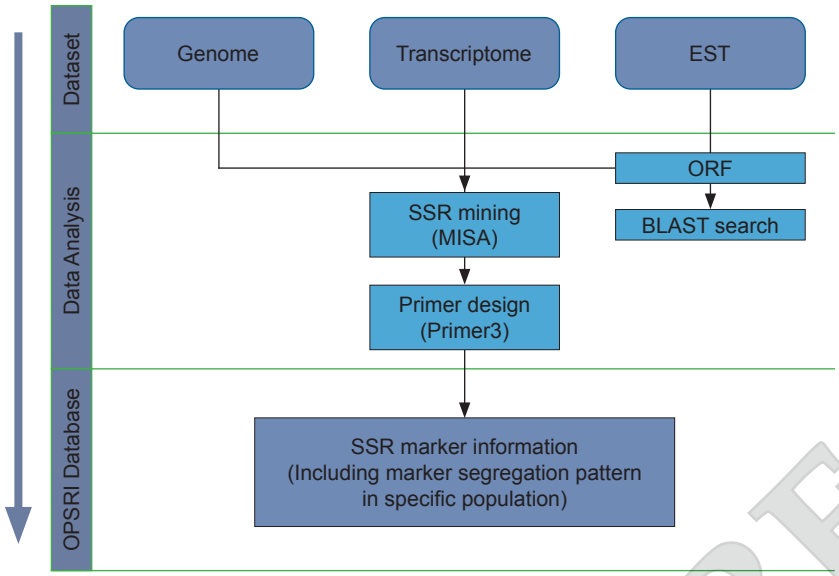

Figure 2. OPSRI analysis pipelines workflow.

RESULTS AND DISCUSSION

The Organisation of OPSRI

OPSRI comprises three main modules - database, analysis and query (*Figure 3*). The OPSRI homepage is divided into three panels, namely the top, navigation and results panels. The top panel includes the Home, Sitemap, Change Password, Logout and Query Search (*Figures 4a-i*). The navigation panel (bottom left panel) contains File Manager, Analysis Tools, Analysis Pipeline, SSR Database, Contact Us and Manual (*Figures 4a-ii*). The results panel (bottom right panel) displays results of the queries or analysis (*Figures 4a-iii*). Users can perform the data query in SSR database navigation panel. In a single webpage, the user can analyse, explore and perform the data query or view the results of a search. The list of publications associated with the data is also shared with users in a graphic link form. A manual pocket book, which is accessible after login, provides detailed instructions to guide users on the analysis module and also to help them get familiar with the system. To facilitate easy and fast interaction, the module "Contact Us" is available to provide support to users. User feedback is included as part of the efforts to improve the system.

Database Features

The OPSRI database at present contains information on a total of 1983 published SSR markers (Ting *et al.*, 2016). These SSRs were developed from various genomic, transcriptome and EST libraries as well as from selective regions in the published oil palm EG5 genome build (Low *et al.*, 2018). The SSR markers from genomic libraries are given the nomenclature sEG, sMg, sMo, sMh, sPSc, sOleiSc and PA while those from transcriptomes are sTEg and those from ESTs are labelled as sEg. The SSRs were utilised for construction of genetic linkage maps for several mapping populations, coded as P2 [Ulu Remis Deli *dura* (ENL48) x Yangambi *pisifera* (ML161)], T128 (self-crossing of a Nigerian *tenera*, T128) and OxG [Colombian *E. oleifera* (UP1026) x Nigerian *tenera* (T128)]. A total of 191, 92 and 156 SSR markers were polymorphic in the three mapping populations, namely P2, T128 and OG, respectively as summarised in *Table 1*. The results showed that a high proportion of the polymorphic SSRs were di(p2)- repeats at 64%, followed by tri(p3)- repeat motifs at 15%. These two repeat motifs were also found to be the most abundant in the oil palm genome (Babu *et al.*, 2019). The identification and recording in the database, of SSR markers that are polymorphic in a specific family is useful to researchers, as they can prioritise these SSR markers for the genetic analysis of specific populations. The utilisation of these SSR markers will also facilitate comparison across different studies especially if they are linked to the same trait in a QTL analysis, which will add confidence to the marker-trait association observed. These are among the main advantages of the present database.

OPSRI Usage

A number of studies have reported on the availability of bioinformatics pipelines for mining of SSRs from sequencing data, examples of which include a system packaged in the tool Galaxy (Griffiths *et al.*, 2016) and a system known as

Figure 3. Overview of OPSRI.

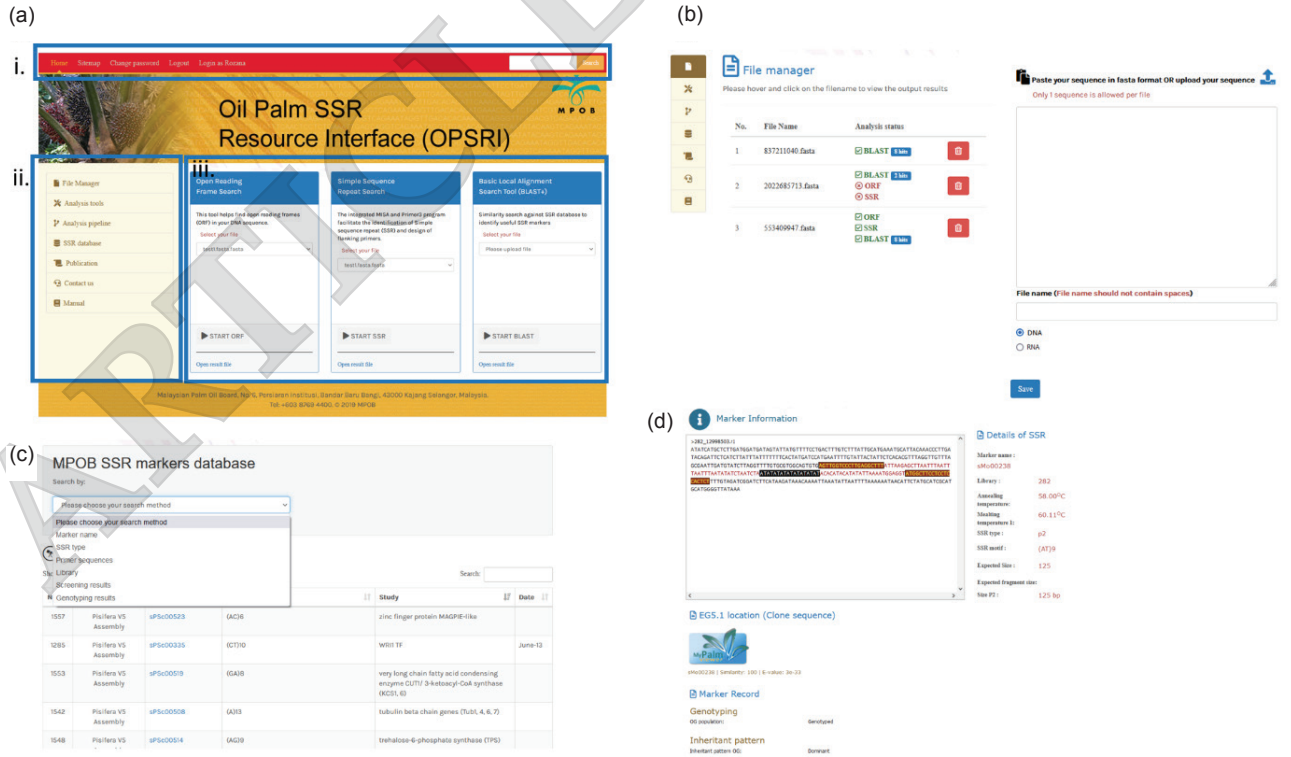

Figure 4. OPSRI Homepage (a) Three panels of OPSRI homepage (b) File Manager: data and results management page (c) Query interface: search SSR database by marker name, SSR type, Primer sequences, Library, Screening and Genotyping results or using a keywords search (d) example of Marker Information page.

TABLE 1. SUMMARY OF SSR MARKERS GENOTYPED IN THREE MAPPING POPULATIONS

No.	SSR type	Library	Polymorphic SSR		
			P2	T128	OxG
1	Compound (C)	EST	1	1	2
		Pisifera V3*	1	-	-
		Pisifera V4*	4	-	2
		Pisifera V5*	2	-	3
		233§ (*Dura*_MF)+	5	-	2
		282§ (Columbian *Oleifera*_MF)+	4	1	3
		Total	17	2	12
2	Imperfect compound (C*)	EST	-	1	1
		Pisifera V3*	1	1	1
		233§ (*Dura*_MF)+	8	2	5
		282§ (Columbian *Oleifera*_MF)+	4	1	1
		Total	13	5	8
3	Mono-repeats (p1)	EST	-	2	2
		Transcriptome	1	1	1
		233§ (*Dura*_MF)+	1	-	1
		280§ (*Dura*_MF)+	1	-	-
		282§ (Columbian *Oleifera*_MF)+	-	-	1
		Total	3	3	5
4	Di-repeats (p2)	EST	8	9	12
		Transcriptome	-	2	2
		Pisifera V1*	6	8	8
		Pisifera V3*	8	6	4
		Pisifera V4*	-	-	2
		Pisifera V5*	5	-	3
		213§ (*Dura*_UF)#	1	-	-
		233§ (*Dura*_MF)+	24	8	20
		280§ (*Dura*_MF)+	1	-	-
		281§ (*Pisifera*_MF)+	31	13	28
		282§ (Columbian *Oleifera*_MF)+	37	12	20
		302§ (Columbian *Oleifera*_MF)+	1	1	1
		Total	122	59	100
5	Tri-repeats (p3)	EST	5	9	9
		Oleifera	-	2	2
		Pisifera V1*	2	3	4
		Pisifera V3*	1	1	1
		Pisifera V4*	2	-	2
		Pisifera V5*	2	-	-
		233§ (*Dura*_MF)+	1	1	1
		281§ (*Pisifera*_MF)+	1	-	-
		282§ (Columbian *Oleifera*_MF)+	10	1	6
		Total	24	17	25
6	Tetra-repeats (p4)	EST	2	1	1
		Pisifera V1*	1	2	1
		234§ (*Dura*_UF)#	1	-	1
		282§ (Columbian *Oleifera*_MF)+	2	-	1
		Total	6	3	4
7	Penta-repeats (p5)	*Pisifera* V3*	2	2	-
		233§ (*Dura*_MF)+	1	-	-
		282§ (Columbian *Oleifera*_MF)+	2	-	-
		Total	5	2	0
8	Hexa-repeats (p6)	233§ (*Dura*_MF)+	-	1	1
		282§ (Columbian *Oleifera*_MF)+	1	-	1
		Total	1	1	2

Note: **Pisifera* V1, V2, V3, V4 and V5 denotes the assembly version of the *Pisifera* genome. #UF standard (no methylation filter) genomic library, +MF methyl filtered genomic library, § Information for the library is available in Low *et al*. (2014), - no polymorphic markers genotyped.

ESAP Plus (Ponyared *et al.*, 2016). More recently, an approach was introduced which can directly identify polymorphic SSRs. The system known as IDSSR, facilitates mining of polymorphic SSRs within regions that were pre-classified as insertion/deletions in a single sequenced genome (Guang *et al.*, 2019). However, apart from the fact that IDSSR is a Perl script-based pipeline which requires some level of computational expertise to execute, the SSRs are mined only from specific regions of the genome, resulting in potential loss of informative markers in other parts of the genome. In addition, SSRome, a web-based database that mined SSRs from 6533 different organisms and further classified them as either originating from genic/non-genic regions among other characteristics, was also reported (Mokhtar and Atia, 2019). However, in the databases described above, experimental details of the SSR markers are not included. In this respect, the main advantage of the OPSRI web-based pipeline reported here compared to the other systems is that it is integrated with the oil palm SSR genotyping database, which facilitates identification of the most informative markers, based on actual experimental data obtained from diverse genetic backgrounds, namely the P2, T128 and OxG families.

Generally, the OPSRI pipelines developed in this study allow for a wider application, ranging from SSR search to primer design (combinations of MISA, Primer3, ORF search script and BLAST). In addition, users can directly search and compare the primers designed with those available within the OPSRI database, to avoid redundancy in assaying for a particular SSR locus. This is a clear advantage over other web-based tools that only focus on specific steps in the analysis *e.g.*, MISA-web which was designed to only search for SSRs (Beier *et al.*, 2017). The OPSRI analysis module also provides two options to users, who can either key in the fasta format sequence into the query box (under File Manager) or upload the sequence file (*Figure 4b*). In OPSRI, users also have the option to analyse the sequence(s) either by selecting a specific programme or using the analysis pipeline linking all the four programmes *i.e.*, MISA, Primer3, ORF search script and BLAST. Results generated from the pipeline can be viewed (via hover or by clicking on the file name) or downloaded.

The query module in the database has been designed with graphical user interface to facilitate data archiving (*Figure 4c*). This is particularly helpful to users who are not familiar with Structural Query Language (SQL) scripting. Search for marker information such as SSR type, motif, primer information and marker genotype profile can be performed by selecting any query option such as marker name (*e.g.*, sMg00026), SSR type (*e.g.*, p3), primer sequence or library (*e.g.*, EST) (*Figure 4d*). Additionally, a free text search for SSR markers linked to specific genes such as the shell and fruit colour genes is also available. The database allows easy retrieval of information related to these markers.

Information available on the motif types and the experimental data facilitates identification of polymorphic SSR primers. Therefore, markers informative in one population can be prioritised to analyse oil palm derived from different genetic backgrounds, enabling a wide range of genetic studies, such as the saturation of genetic linkage maps. Users can register and access the OPSRI database at no cost at http://opsri.mpob.gov.my. Standard users will be able to view and download publicly available SSR primers, while OPSRI pre-registered users at MPOB will be able to access both public and private databases.

Distribution of Repeat Motifs in Oil Palm EST to Enrich the Information in OPSRI

In order to enhance the information available in OPSRI, a subset of EST data from a NCBI Genbank was mined for SSR markers and the positions of the repeat motifs were either located in the UTR or CDS regions. EST cluster analysis revealed 13 600 consensus and 7631 singletons giving a total of 21 231 unique sequences. Twenty sequences with fewer than 100 nucleotides were excluded from further analysis. The subsequent SSR search showed that 2465 SSRs were found in 2014 EST clones. The 2465 SSRs consisted of 983 (39.9%) mononucleotides, 794 (32.2%) dinucleotides, 650 (26.4%) trinucleotides, 31 (1.25%) tetranucleotides, 4 (0.16%) pentanucleotides and 3 (0.12%) hexanucleotides. The distribution of the number of repeats observed across the SSR motifs is shown in *Table 2*. The high number of dinucleotides observed in the ESTs utilised in this study compared to trinucleotides is consistent with that observed by Babu *et al.*, (2019) in oil palm. Among the dinucleotide repeats, the most common motif is AG/CT (75.5%), whereas CG/CG (0.1%) is the least abundant motif. The AG/CT repeats are also the most copious motifs in many plants (Guo *et al.*, 2017; Liu *et al.*, 2018; Rabeh *et al.*, 2018; Wan *et al.*, 2020). With respect to the tri-nucleotide repeats, AGG/CCT motif was the most common, followed by AAG/CTT and CCG/CGG. The A/T motif was the most abundant repeat motif observed in this study. Tóth *et al.* (2000) also observed an abundance of A/T motifs in various eukaryotic genomes, which the authors reported was likely due to the poly (A/T) tails of specific retrotransposon's (especially ALU and LINE-1) scattered across the genome. Interestingly the number of A/T repeats observed declined with increase in repeat length (*Table 1*) as was observed for monocotyledons by Qin *et al.* (2015).

TABLE 2. FREQUENCY AND DISTRIBUTION OF THE DIFFERENT SSR TYPES IDENTIFIED IN 2014 OIL PALM ESTs

Repeats	No. of repeats												Total
	5	6	7	8	9	10	11	12	13	14	15	>15	
A/T	-	-	-	-	-	215	130	88	86	63	52	240	874
C/G	-	-	-	-	-	40	29	7	11	4	7	11	109
AC/GT	-	7	19	10	12	5	3	1	1	2	1	5	66
AG/CT	-	33	133	110	110	49	34	16	12	15	10	70	592
AT/AT	-	11	21	19	28	14	9	9	3	2	2	14	132
CG/CG	-	3	-	1	-	-	-	-	-	-	-	-	4
AAC/GTT	8	6	2	2	-	1	-	-	-	-	-	-	19
AAG/CTT	55	29	17	9	6	1	2	-	-	-	-	1	120
AAT/ATT	15	8	5	-	4	2	1	2	-	-	-	1	38
ACC/GGT	38	17	4	-	1	1	-	1	-	1	-	-	63
ACG/CGT	7	1	3	2	-	-	-	-	-	-	-	-	13
ACT/AGT	2	-	1	-	-	-	-	-	1	-	1	-	5
AGC/CTG	36	21	11	11	1	2	-	-	-	-	-	1	83
AGG/CCT	78	32	12	8	6	1	-	-	-	-	-	-	137
ATC/ATG	34	14	5	3	3	2	-	-	-	-	-	-	61
CCG/CGG	65	24	13	4	4	1	-	-	-	-	-	-	111
AAAG/CTTT	2	1	-	-	-	-	-	-	-	-	-	-	3
AAAT/ATTT	6	1	-	-	-	-	-	-	-	-	-	-	7
AACC/GGTT	-	1	-	-	-	-	-	-	-	-	-	-	1
AAGC/CTTG	-	1	-	-	-	-	-	-	-	-	-	-	1
AAGG/CCTT	-	-	1	-	-	-	-	-	-	-	-	-	1
ACAT/ATGT	4	2	-	-	-	1	-	-	-	-	-	-	7
ACGC/CGTG	1	-	-	-	-	-	-	-	-	-	-	-	1
ACGT/ACGT	1	-	-	-	-	-	-	-	-	-	-	-	1
AGAT/ATCT	-	-	-	-	1	-	-	-	-	-	-	-	1
AGCG/CGCT	1	-	-	-	-	-	-	-	-	-	-	-	1
AGGC/CCTG	1	-	-	-	-	-	-	-	-	-	-	-	1
AGGG/CCCT	1	1	-	-	-	-	-	-	-	-	-	-	2
ATGC/ATGC	2	1	-	-	-	-	-	-	-	-	-	-	3
CCCG/CGGG	1	-	-	-	-	-	-	-	-	-	-	-	1
AAAAT/ATTTT	2	-	-	-	-	-	-	-	-	-	-	-	2
AAGGG/CCCTT	1	-	-	-	-	-	-	-	-	-	-	-	1
AATAT/ATATT	1	-	-	-	-	-	-	-	-	-	-	-	1
ACCTCG/AGGTCG	1	-	-	-	-	-	-	-	-	-	-	-	1
ACGCCG/CGGCGT	1	-	-	-	-	-	-	-	-	-	-	-	1
AGCCTG/AGGCTC	-	1	-	-	-	-	-	-	-	-	-	-	1
N													983
NN													794
NNN													650
NNNN													31
NNNNN													4
NNNNNN													3

Note: - No SSR identified.

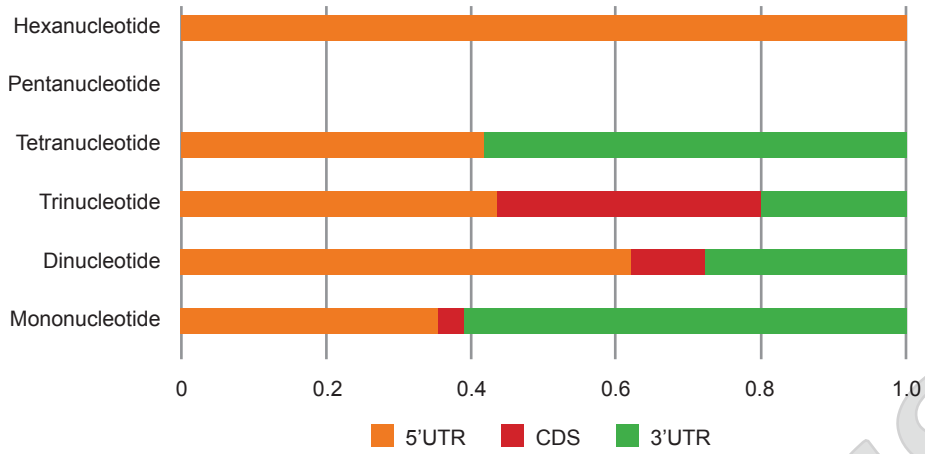

Figure 5. A comparison of the distribution of different SSR repeat motifs in the 5'UTR, 3'UTR and CDS.

The location of each SSR, whether in the UTR or coding region, was identified and validated through the Blast programme. The results revealed 481 ESTs with 544 SSRs in the 5'UTR, 316 ESTs with 372 SSRs in the 3'UTR and 150 ESTs with 175 SSRs in the coding region. An analysis of the SSR motifs in the 5'UTR, 3'UTR and coding regions revealed that 84, 70 and 14 ESTs, respectively have more than one SSR. The analysis also showed that the UTRs contained more SSRs compared to coding regions, similar to the pattern observed in rice and Arabidopsis (Lawson *et al.*, 2006). The distribution of different repeat type classes also showed that the 5'UTR had higher dimeric repeats (247/397) and trinucleotides (144/331), while the 3'UTR showed more mono repeats (188/350). In CDS, the large number of trinucleotide motifs (121/331) are probably involved in increasing the protein size due to the repeat domains. For the dinucleotide repeats, there was a bias towards the AG/CT motif in both UTRs, with 221 found in the 5'UTR and 67 in the 3'UTR. Similarly, in the CDS, 37 dinucleotide repeats contained the AG/CT motif. The abundance of the dinucleotide AG/CT SSR motifs was also reported in other oil plants such as *Arachis hypogaea* (Wang *et al.*, 2017) and *Elaeagnus mollis* (Liu *et al.*, 2020). Interestingly, the 5' UTR contained more trinucleotides than the 3' UTR (44% *vs*. 20% SSR). *Figure 5* shows a comparison of the distribution of the different SSR repeat motifs in the 5'UTR, 3'UTR and CDS.

These observations have been incorporated into the database and may assist in selecting and designing primers flanking the appropriate motifs for experimental purposes. Selecting dinucleotide motifs in UTR and trinucleotides in CDS, especially those with a long repeat type may improve chances of identifying polymorphic SSR markers in oil palm populations. However, it is important to note that the association between the polymorphism level of an SSR and the length of SSR motif is species dependent (Hou *et al.*, 2017; Sigang *et al.*, 2021), and as such, when the information becomes available for oil palm it will be updated on the database. For example, polymorphism rates were not associated with the length of the SSR in peanut, where the polymorphism decreased as motif repeat number increased (Zhao *et al.*, 2012). In addition, Scott *et al.* (2000) observed that in grapes, the level of polymorphism for the SSR uncovered from the three different regions of an EST varied at the taxonomy level (genera, cultivar and species). The study further revealed that SSRs from the 3'UTR were most polymorphic when screening cultivars, while SSRs at the 5'UTR were informative when comparing cultivar of different species. The SSRs in CDS were most polymorphic between species and samples from related genera. Thus, this and other relevant information can serve as a guide to researchers to select and prioritise EST-SSR markers from the OPSRI database for use in their own research programmes.

CONCLUSION

In this study, the development of a database integrated with bioinformatics tools (MISA, Primer3, ORF search script and BLAST), facilitates the mining of SSRs from a collection of oil palm sequences. The database is further enriched with experimental data, and information on frequency and distribution of repeat motifs as well as location of the SSR motifs in a gene. Such information is highly useful for selecting candidate SSR markers for use in research. As such, the systematic archiving of information on the SSRs markers in this web-based tool, should prove useful for oil palm genetic studies. More importantly, researchers will also have the opportunity to select SSR markers-

based on their previous performance (especially polymorphism rate in specific families), length of repeat motifs or position in the genome, which will accelerate implementation of genomics guided breeding programmes in oil palm. The database facilitates data-sharing among researchers, where the benefit extends beyond oil palm to other plants from closely related genera and taxa such as coconut and date palm.

ACKNOWLEDGEMENT

We thank the Director-General of MPOB for permission to publish this article. We would also like to acknowledge colleagues of the Bioinformatics and Genomics Units for their kind assistance. Special thanks to Prof. Denis Murphy from the University of South Wales, Dr. Chan Pek Lan and Dr. S. Ravigadevi for valuable comments on the article.

REFERENCES

Al-Faifi, S A; Migdadi, H M; Algamdi, S S; Khan, M A; Ammar, M H; Al-Obeed, R S; Al-Thamra, M I; El-Harty, E H and Jakse, J (2016). Development, characterization and use of genomic SSR markers for assessment of genetic diversity in some Saudi date palm (*Phoenix dactylifera* L.) cultivars. *Electron. J. Biotechnol.*, *21*: 18-25.

Altschul, S F; Madden, T L; Schäffer, A A; Zhang, J; Zhang, Z; Miller, W and Lipmann, D J (1997). Gapped BLAST and PSI- BLAST: A new generation of protein database search programs. *Nucleic Acids Res.*, *25(17)*: 3389-3402.

Babu, K; Mary Rani, K L; Sahu, S; Mathur, R K; Naveen Kumar, P; Ravichandran, G; Anitha, P and Bhagya, H P (2019). Development and validation of whole genome-wide and genic microsatellite markers in oil palm (*Elaeis guineensis* Jacq.): First microsatellite database (OpSatdb). *Sci. Rep.*, *9(1)*: 1899.

Backiyarani, S; Uma, S; Varatharj, P and Saraswathi, M S (2013). Mining of EST-SSR markers of *Musa* and their transferability studies among the members of order the Zingiberales. *Appl. Biochem. Biotechnol.*, *169(1)*: 228-238.

Bagshaw, A T M (2017). Functional mechanisms of microsatellite DNA in eukaryotic genomes. *Genome Biol. Evol.*, *9(9)*: 2428-2443.

Bazzo, B R; de Carvalho, L M; Carazzolle, M F; Pereira, G A G and Colombo, C A (2018). Development of novel EST-SSR markers in the macaúba palm (*Acrocomia aculeata*) using transcriptome sequencing and cross-species transferability in Arecaceae species. *BMC Plant Biol.*, *18(1)*: 276.

Beier, S; Thiel, T; Münch, T; Scholz, U and Mascher, M (2017). MISA-web: A web server for microsatellite prediction. *Bioinformatics*, *33(16)*: 2583-2585.

Bhagya, H P; Kalyana Babu, B; Gangadharappa, P M; Naika, M B N; Satish, D and Mathur, R K (2020). Identification of QTLs in oil palm (*Elaeis guineensis* Jacq.) using SSR markers through association mapping. *J. Genet.*, *99(1)*: 19.

Bhattarai, G; Shi, A; Kandel, D R; Solís-Gracia, N; da Silva, J A and Avila, C A (2021). Genome-wide simple sequence repeats (SSR) markers discovered from whole-genome sequence comparisons of multiple spinach accessions. *Sci. Rep.*, *11(1)*: 9999.

Du, L; Zhang, C; Liu, Q; Zhang, X and Yue, B (2018). Krait: An ultrafast tool for genome-wide survey of microsatellites and primer design. *Bioinformatics*, *34(4)*: 681-683.

Fan, M; Gao, Y; Gao, Y; Wu, Z; Liu, H and Zhang, Q (2019). Characterization and development of EST-SSR markers from transcriptome sequences of chrysanthemum (*Chrysanthemum* × *morifolium* Ramat.). *HortScience*, *54(5)*: 772-778.

Gou, X; Shi, H; Yu, S; Wang, Z; Li, C; Liu, S; Ma, J; Chen, G; Liu, T and Liu, Y (2020). SSRMMD: A rapid and accurate algorithm for mining SSR feature loci and candidate polymorphic SSRs based on assembled sequences. *Front. Genet.*, *11*: 706.

Griffiths, S M; Fox, G; Briggs, P J; Donaldson, I J; Hood, S; Richardson, P; Leaver, G W; Truelove, N K and Preziosi, R F (2016). A galaxy-based bioinformatics pipeline for optimised, streamlined microsatellite development from Illumina next-generation sequencing data. *Conservation Genet. Resour.*, *8(4)*: 481-486.

Guang, X M; Xia, J Q; Lin, J Q; Yu, J; Wan, Q H and Fang, S G (2019). IDSSR: An efficient pipeline for identifying polymorphic microsatellites from a single genome sequence. *Int. J. Mol. Sci.*, *20(14)*: 3497.

Guo, R; Landis, J B; Moore, M J; Meng, A; Jian, S; Yao, X and Wang, H (2017). Development and application of transcriptome-derived microsatellites in *Actinidia eriantha* (Actinidiaceae). *Front. Plant Sci.*, *8*: 1383.

Gupta, P K; Varshney, R K; Sharma, P C and Ramesh, B (1999). Molecular markers and their applications in wheat breeding. *Plant Breeding*, *118(5)*: 369-390.

Hamelin, C; Sempere, G; Jouffe, V and Ruiz, M (2012). TropGeneDB, the multi-tropical crop information system updated and extended. *Nucleic Acids Res.*, *41(D1)*: D1172-D1175.

Hou, B; Feng, S and Wu, Y (2017). Systemic identification of *Hevea brasiliensis* EST-SSR markers and primer screening. *J. Nucleic Acids*, *2017*: 1-9.

Huang, X and Madan, A (1999). CAP3: A DNA sequence assembly program. *Genome Res.*, *9(9)*: 868-877.

Kushairi, A; Mohd Din, A and Rajananidu, N (2011). Oil palm breeding and seed production. *Further Advances in Oil Palm Research (2000-2010)* (Mohd Basri, W; Choo, Y M and Chan, K W eds.). Vol. 1. MPOB, Bangi. p. 47-101.

Lawson, M J and Zhang, L (2006). Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. *Genome Biol.*, *7(2)*: 1-11.

Lebedev, V G; Subbotina, N M; Maluchenko, O P; Lebedeva, T N; Krutovsky, K V and Shestibratov, K A (2020). Transferability and polymorphism of SSR markers located in flavonoid pathway genes in *Fragaria* and *Rubus* species. *Genes*, *11(1)*: 11.

Li, Y C; Korol, A B; Fahima, T and Nevo, E (2004). Microsatellites within genes: Structure, function and evolution. *Mol. Biol. Evol.*, *21(6)*: 991-1007.

Liu, Y; Guo, Y; Xing, D and Long, C (2018). Development and characterization of genomic simple sequence repeats for *Colocasia gigantea* (Blume) Schott using 454 sequencing. *Chilean J. Agric. Res.*, *78(1)*: 23-29.

Liu, Y; Li, S; Wang, Y; Liu, P and Han, W (2020). De *novo* assembly of the seed transcriptome and search for potential EST-SSR markers for an endangered, economically important tree species: *Elaeagnus mollis* diels. *J. For. Res.*, *31(3)*: 759-767.

Low, E-T L; Rosli, R; Jayanthi, N; Mohd-Amin, A H; Azizi, N; Chan, K L; Maqbool, N J; Maclean, P; Brauning, R; McCulloch, A; Moraga, R; Ong-Abdullah, M and Singh, R (2014). Analyses of hypomethylated oil palm gene space. *PLoS ONE*, *9(1)*: p.e86728.

Low, E-T L; Jayanthi, N; Chan, K-L; Sanusi, N S N M; Ab Halim, M A; Rosli, R; Azizi, N; Amiruddin, N; Angel, L P L; Ong-Abdullah, M; Singh, R; Manaf, M A A; Sambanthamurthi R; Parveez, G K A and Kushairi, A (2018). The oil palm genome revolution. *J. Oil Palm Res.*, *29(4)*: 456-468.

Low, E-T L; Azizi, N; Halim, M A A; Sanusi, N S N M; Chan, K-L; Amiruddin, N; Jayanthi, N; Ong-Abdullah, M; Singh, R; Sambanthamurthi, R; Manaf, M A A A and Kushairi, A (2020). Oil palm genome: Strategies and applications. *The Oil Palm Genome* (Ithnin, M and Din, A K eds.). *Compendium of Plant Genomes*. Springer, Cham. p. 83-115.

Metz, S; Cabrera, J M; Rueda, E; Giri, F and Amavet, P (2016). FullSSR: Microsatellite finder and primer designer. *Adv. Bioinform.*, *2016*: 1-4.

Mokhtar, M M and Atia, M A M (2019). SSRome: An integrated database and pipelines for exploring microsatellites in all organisms. *Nucleic Acids Res.*, *47(D1)*: D244-D252.

O'Leary, N A; Wright, M W; Brister, J R; Ciufo, S; Haddad, D; McVeigh, R; Rajput, B; Robbertse, B; Smith-White, B; Ako-Adjei, D; Astashyn, A; Badretdin, A; Bao, Y; Blinkova, O; Brover, V; Chetvernin, V; Choi, J; Cox, E; Ermolaeva, O; Farrell, C M; Goldfarb, T; Gupta, T; Haft, D; Hatcher, E; Hlavina, W; Joardar, V S; Kodali, V K; Li, W; Maglott, D; Masterson, P; McGarvey, K M; Murphy, M R; O'Neill, K; Pujar, S; Rangwala, S H; Rausch, D; Riddick, L D; Schoch, C; Shkeda, A; Storz, S S; Sun H; Thibaud-Nissen, F; Tolstoy, I; Tully, R E; Vatsan, A R; Wallin, C; Webb, D; Wu, W; Landrum, M J; Kimchi, A; Tatusova, T; DiCuccio, M; Kitts, P; Murphy, T D and Pruitt, K D (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, *44(D1)*: D733-D745.

Ponyared, P; Ponsawat, J; Tongsima, S; Seresangtakul, P; Akkasaeng, C and Tantisuwichwong, N (2016). ESAP plus: A web-based server for EST-SSR marker development. *BMC Genomics*, *17(13)*: 163-173.

Qin, Z; Wang, Y; Wang, Q; Li, A; Hou, F and Zhang, L (2015). Evolution analysis of simple sequence repeats in plant genome. *PLoS ONE*, *10(12)*: p.e0144108.

Rabeh, K; Gaboun, F; Belkadi, B and Filali-Maltouf, A (2018). *In silico* development of new SSRs primer for aquaporin linked to drought tolerance in plants. *Plant Signal. Behav.*, *13(11)*: 1-7.

Rozen, S and Skaletsky, H (2000). Primer3 on the WWW for general users and for biologist programmers. *Bioinformatics Methods Protocols* (Misener, S and Krawetz, S A eds.). *Methods in Molecular Biology™. Vol. 132*. Humana Press, Totowa, NJ. p. 365-386.

Sarimana, U; Herrero, J; Erika, P; Indarto, N; Wendra, F; Santika, B; Ritter, E; Sembiring, Z and Asmono, D (2021). Analysis of genetic diversity and discrimination of oil palm DxP populations based on the origins of *pisifera* elite parents. *Breeding Sci.*, *71(2)*: 134-143.

Scott, K D; Eggler, P; Seaton, G; Rossetto, M; Ablett, E M; Lee, L S and Henry, R J (2000). Analysis of SSRs derived from grape ESTs. *Theor. Appl. Genet.*, *100*: 723-726.

Sigang, F; Hao, H; Yong, L; Pengfei, W; Chao, Z; Lulu, Y; Xiuting, Q and Qiu, L (2021). Genome-wide identification of microsatellite and development of polymorphic SSR markers for spotted sea bass (*Lateolabrax maculatus*). *Aquac. Rep.*, *20*: 100677.

Singh, R; Nagappan, J; Tan, S G; Panandam, J M and Cheah, S C (2007). Development of simple sequence repeat (SSR) markers for oil palm and their application in genetic mapping and fingerprinting of tissue culture clones. *Asia Pac. J. Mol. Biol. Biotechnol.*, *15(3)*: 121-131.

Singh, R; Tan, S G; Panandam J M; Rahimah, A R; Leslie Ooi, L C L; Low, E T L; Sharma, M; Jansen, J and Cheah, S C (2009). Mapping quantitative trait loci (QTLs) for fatty acid composition in an interspecific cross of oil palm. *BMC Plant Biol.*, *9*: 114.

Singh, R; Ong-Abdullah, M; Low, E T L; Abdul Manaf, M A; Rosli, R; Nookiah, R; Leslie Ooi, C L; Ooi, S-E; Chan, K-L; Halim, M A; Azizi, N; Nagappan, J; Bacher, B; Lakey, N; Smith, S W, He, D; Hogan, M; Budiman, M A; Lee, E K; DeSalle, R; Kudrna, D; Goicoechea, J L; Wing, R A; Wilson, R K; Fulton, R S; Ordway, J M; Martienssen, R A and Sambanthamurthi, R (2013). Oil palm genome sequence reveals divergence of interfertile species in old and new worlds. *Nature*, *500(7462)*: 335-339.

Song, X Y; Zhang, C Z; Li, Y; Feng, S S; Yang, Q and Huang, S W (2016). SSR analysis of genetic diversity among 192 diploid potato cultivars. *Hortic. Plant J.*, *2(3)*: 163-171.

Sorkheh, K; Prudencio, A S; Ghebinejad, A; Dehkordi, M K; Erogul, D; Rubio, M and Martínez-Gómez, P (2016). *In silico* search, characterization and validation of new EST-SSR markers in the genus *Prunus. BMC Res. Notes*, *9(1)*: 1-11.

Soto-Cerda, B J; Saavedra, H U; Navarro, C N and Ortega, P M (2011). Characterization of novel genic SSR markers in *Linum usitatissimum* (L.) and their transferability across eleven *Linum* species. *Electron. J. Biotechnol.*, *14(2)*: 4-4.

Sunilkumar, K; Murugesan, P; Mathur, R K and Rajesh, M K (2020). Genetic diversity in oil palm (*Elaeis guineensis* and *Elaeis oleifera*) germplasm as revealed by microsatellite (SSR) markers. *Indian J. Agric. Sci.*, *90(4)*: 69-73.

Tautz, D and Renz, M (1984). Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Res.*, *12(10)*: 4127-4138.

Thiel, T; Michalek, W; Varshney, R and Graner, A (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Gen.*, *106(3)*: 411-422.

Ting, N C; Jansen, J; Nagappan, J; Ishak, Z; Chin, C W; Tan, S G; Cheah, S C and Singh, R (2013). Identification of QTLs associated with callogenesis and embryogenesis in oil palm using genetic linkage maps improved with SSR markers. *PLoS ONE*, *8(1)*: p.e53076.

Ting, N C; Jansen, J; Mayes, S; Massawe, F; Sambanthamurthi, R; Ooi, L C L; Chin, C W; Arulandoo, X; Seng, T Y; Alwee, S S R S and Ithnin, M (2014). High density SNP and SSR-based genetic maps of two independent oil palm hybrids. *BMC Genom.*, *15(1)*: 1-11.

Ting, N C; Yaakub, Z; Kamaruddin, K; Mayes, S; Massawe, F; Sambanthamurthi, R; Jansen, J; Low, L E T; Ithnin, M; Kushairi, A; Arulandoo, X; Rosli, R; Chan, K L; Amiruddin, N; Sritharan, K; Lim, C C; Nookiah, R; Amiruddin, M D and Singh, R (2016). Fine-mapping and cross-validation of QTLs linked to fatty acid composition in multiple independent interspecific crosses of oil palm. *BMC Genom.*, *17(1)*: 1-17.

Tóth, G; Gáspári, Z and Jurka, J (2000). Microsatellites in different eukaryotic genomes: Survey and analysis. *Genome Res.*, *10(7)*: 967-981.

Tranbarger, T J; Kluabmongkol, W; Sangsrakru, D; Morcillo, F; Tregear, W J; Tragoonrung, S and Billotte, N (2012). SSR markers in transcripts of genes linked to post-transcriptional and transcriptional regulatory functions during vegetative and reproductive development of *Elaeis guineensis*. *BMC Plant Biol.*, *12(1)*: 1-12.

Varshney, R K; Graner, A and Sorrells, M E (2005). Genic microsatellite markers in plants: Features and applications. *Trends Biotechnol.*, *23(1)*: 48-55.

Vieira, M L C; Santini, L; Diniz, A L and Munhoz, C D F (2016). Microsatellite markers: What they mean and why they are so useful. *Genet. Mol. Biol.*, *39(3)*: 312-328.

Wan, Y; Zhang, M; Hong, A; Zhang, Y and Liu, Y (2020). Characteristics of microsatellites mined from transcriptome data and the development of novel markers in *Paeonia lactiflora*. *Genes*, *11(2)*: 214.

Wang, X and Wang, L (2016). GMATA: An integrated software package for genome-scale SSR mining, marker development and viewing. *Front. Plant Sci.*, *7*: 1350.

Wang, H; Lei, Y; Yan, L; Wan, L; Cai, Y; Yang, Z; Lv, J; Zhang, X; Xu, C and Liao, B (2017). Development and validation of simple sequence repeat markers from *Arachis hypogaea* transcript sequences. *Crop J.*, *6(2)*: 172-180.

Zaki, N M; Singh, R; Rosli, R and Ismail, I (2012). *Elaeis oleifera* genomic-SSR markers: Exploitation in oil palm germplasm diversity and cross-amplification in Arecaceae. *Int. J. Mol. Sci.*, *13(4)*: 4069-4088.

Zhao, Y; Prakash, C S and He, G (2012). Characterization and compilation of polymorphic simple sequence repeat (SSR) markers of peanut from public database. *BMC Res. Notes*, *5(1)*: 1-7.

Zolkafli, S H; Ithnin, M; Chan, K-L; Zainol Abidin, M I; Ismail, I; Ting, N C; Ooi, L C-L and Singh, R (2021). Optimal set of microsatellite markers required to detect illegitimate progenies in selected oil palm (*Elaeis guineensis* Jacq.) breeding crosses. *Breeding Sci.*, *71(2)*: 253-260.