Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Genetic diversity analysis of Pandanus amaryllifolius using EST-SSR and iPBS markers

Norhafizah Sidek
Universiti Malaya
Boon Chin Tan
Universiti Malaya
Katharina Mebus
Universiti Malaya
Sima Taheri
Universiti Malaya
Setiawan Agus Budi
Universitas Gadjah Mada
Chee How Teo
cheehow.teo@um.edu.my
Universiti Malaya

Research Article

Keywords: Pandanus amaryllifolius, Expressed sequence tag (EST)-simple sequence repeats (SSR), RNA-seq, inter-primer binding site (iPBS)

Posted Date: December 22nd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3779650/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations:
No competing interests reported.

Tables are available in the Supplementary Files section.

Genetic diversity analysis of *Pandanus amaryllifolius* using EST-SSR and iPBS markers

Abstract

Pandanus amaryllifolius, a tropical fragrant screw pine, is commonly used in cooking and traditional medicine. Despite the fact that various studies have been conducted on the metabolite, transcriptome and proteome profiles of *P. amaryllifolius*, there is a scarcity of molecular markers suitable for genetic improvement. Therefore, this research aimed to analyse the genetic diversity of *P. amaryllifolius* using EST-SSR markers derived from transcriptome dataset and iPBS marker system. Using the leaf transcriptomes of three biological replicates, we successfully generated 157,467 unigenes from *P. amaryllifolius,* with an average length of 1,084 base pairs. Of these, 66,820 EST-SSRs were identified, presenting one SSR for every 2.6 kb of distribution density. The most common SSRs were dinucleotides (45.54%), followed by mononucleotides (32.65%). Out of 48,816 developed EST-SSR markers, we randomly selected 30 for the genetic diversity analysis among 24 *P. amaryllifolius* accessions, together with sixteen iPBS markers for *P. amaryllifolius* diversity study. Only two EST-SSR and three iPBS markers showed polymorphic bands, indicating a low polymorphism level among 24 *P. amaryllifolius* accessions. Diversity analysis using EST-SSR markers revealed 3 polymorphic bands with an average polymorphic information content (PIC) value of 0.26. Meanwhile, the iPBS markers generated10 polymorphic bands with an average PIC value of 0.16. The UPGMA cluster analysis differentiated accessions into 5 clusters with iPBS markers and 2 clusters with EST-SSR markers, indicating that iPBS marker system is more effective in identifying the polymorphism of this species. The present work provides a great resource for identifying genes and developing markers in *P. amaryllifolius.*

Keywords

Pandanus amaryllifolius, Expressed sequence tag (EST)-simple sequence repeats (SSR), RNA-seq, inter-primer binding site (iPBS)

Declaration

Funding

This work was supported by the RU grant of Universiti Malaya (RU006-2019) and the MOHE-Top 100 grant (TU002B-2018).

Conflict of interests

The authors declare that they have no competing interests.

Data availability

Transcriptome datasets supporting the conclusions of this article are available in

Code availability

Not applicable

Authors' contributions

SN conducted the experiments. CHT, BCT, and MK conceived the idea and designed the experiments. All authors wrote and approved the manuscript.

Ethics approval

Not applicable

Introduction

Pandanus amaryllifolius (common name: pandan), a tropical plant in the screw pine family Pandanaceae, is commonly found in Southeast Asia, such as Thailand, Malaysia, Indonesia, and India (Mohd Amnan et al. 2023). Similarly to Basmati-India, Jasmine-Thai, and Kaorimai-Japan rice, *P. amaryllifolius* is well-known for its exotic fragrance emitted by its leaves (Wongpornchai 2006). Besides their aromatic value, the leaves are often used in cooking to give flavour and natural colour to foods. 2-acetyl-1-pyrroline (2AP) is the major chemical compound of the leaves (Yahya et al. 2011), while essential oils, carotenoids, tocopherols and tocotrienols also present (Lee et al. 2004).

The leaves of this plant species have been used as a folk medicine to energise the body, relieve indigestion and treat fever (Reshidan et al. 2019) and diabetes mellitus (Kaban and Putri 2020). In addition, some alkaloids from *P. amaryllifolius* leaves exhibited antimicrobial activity and antihyperglycemic effects (Chiabchalard and Nooron 2015; Laluces et al. 2015; Wakte et al. 2007). Although there are reports on the metabolite (Ghasemzadeh and Jaafar 2013), transcriptome (Bhatt et al. 2021) and proteome (Amnan et al. 2022) profiles of *P. amaryllifolius*, there has been little analysis of the genetic variations within *P. amaryllifolius,* hindering its use in genetic and breeding studies. Previously, a study by Wakte et al. (2012) utilised ISSR and amplified fragment length polymorphism (AFLP) markers to evaluate the genetic diversity between *P. amaryllifolius* accessions from India, Sri Lanka and Thailand. However, until now, the genetic diversity of this species in Malaysia has not been evaluated.

Simple sequence repeat (SSR) marker system has been widely utilised to study the genetic basis of important traits and the mapping of regulatory genes in plants (Wang et al. 2018). SSRs are repetitive DNA sequences found in both protein-coding and non-coding regions of DNA sequences. They are abundant, highly polymorphic, multiallelic, codominant, reproducible, and transferable (Boopathi 2020). Nevertheless, generating SSR markers from genomic DNA is both costly and ineffective, resulting in limited transferability of these

markers. Therefore, expressed sequence tag (EST)–SSR markers system is preferable, particularly in organisms that lack of sequence database of their own (Cai et al. 2019). High-throughput NGS technologies, such as RNA sequencing (RNA-seq), offer a new platform to obtain large transcriptome sequences of non-model species. The transcriptome sequences generated from RNA-seq can mine SSR marker sequences and develop the required primers for EST-SSR markers (Taheri et al. 2019; Zhang et al. 2019).

In addition to SSR marker systems, the inter-primer binding site (iPBS) marker has gained attention for analysing genetic variation in plant species. This marker system was developed by Kalendar et al. (2010) and has been effectively used to evaluate the genetic diversity in various plant species (Aydın et al. 2020; Bayat et al. 2018; Borna et al. 2017; Coutinho et al. 2018). The iPBS marker system was developed based on the conserved regions of the primer binding site (PBS) sequences for transfer RNAs (tRNAs), which are used in reverse transcription during the replication cycle of the long terminal repeat (LTR) retrotransposons (Guo et al. 2014). The iPBS is a useful DNA fingerprinting tool without requiring previous genomic DNA information. The iPBS marker system is able to differentiate among closely related genotypes. These markers exhibit a significant level of reproducibility attributed to the length of the primers and the great stringency that the annealing temperature achieves (Amom and Nongdam 2017; Bonchev et al. 2019).

The primary objective of this research was to analyse the genetic diversity of *P. amaryllifolius* utilising a combination of EST-SSR markers derived from the transcriptome dataset and the iPBS marker system. This transcriptome and genetic variation analyses provide new perspectives into the evolutionary origin of *P. amaryllifolius* and are valuable resources for molecular breeding to develop improved cultivars.

Materials and Methods

Plant material and RNA extraction

P. amaryllifolius plants purchased from a nursery in Sungai Buloh, Selangor, Malaysia, were grown in clay silt loam soil and acclimatised in a growth room at the Universiti Malaya, Malaysia. All plants were maintained at 25°C with a cycle of 16 h of light and 8 h of darkness. After two weeks of acclimatisation, young leaves from three *P. amaryllifolius* plants were sampled, immediately frozen in liquid nitrogen, and stored at -80°C. Total RNAs were extracted using a modified CTAB method (Tattersall et al. 2005). The leaf samples were ground into fine powder in liquid nitrogen using a sterilised mortar and pestle. Subsequently, 5 ml preheated extraction buffer was added to 1 g of sample, followed by an incubation step at 65°C for 10 min. The aqueous phase was recovered and subjected to two rounds of chloroform: isoamyl (24:1) alcohol extraction. The upper aqueous phase was recollected and transferred to a new centrifuge tube. After adding 2.5 M of lithium chloride (Calbiochem, USA), the mixture was incubated at -20°C overnight. Next, the samples were spun at 11,000 × *g* for 30 min at 4°C. After washing with 70% ethanol, the pellet was briefly dried before adding 50 µl DEPC-treated water (Simms et al. 1993). The RNA samples were kept at -80°C for future use.

cDNA library construction and RNA sequencing

mRNA from each sample was purified using Oligo (dT) magnetic beads (Thermo Fisher Scientific, USA). The mixture was supplemented with a fragmentation buffer to induce mRNA breakdown before synthesising complementary DNA (cDNA) using reverse transcription-polymerase chain reaction (RT-PCR) and random primers. The short cDNA fragments were purified and resolved using the elution buffer. These fragments were then subjected to end reparation using T4 DNA and Klenow DNA polymerase (Thermo Fisher Scientific, USA). Subsequently, a 3'-terminal adenylation was added to the mixture to repair cDNA fragments

and link with sequencing adapters. A cDNA library was created by utilising purified products and amplified through PCR. Prior to sequencing on an Illumina HiSeq 4000 system at Beijing Genomics Institute (Shenzhen, China), the quality of cDNA was assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies, Germany).

De novo assembly and functional annotation of unigenes

The generated raw reads from sequencing were checked to exclude low-quality reads (>20% Q-score of <20 bases), adaptor-polluted reads, unclear reads with >5% unknown base (N), and non-coding RNA. Then, the 'Clean Reads' were saved in FASTQ format (Cock et al. 2010). The Trinity program (v2.8.5) was used to assemble the clean reads to obtain *P. amaryllifolius* transcriptome dataset (Grabherr et al. 2011). The TGICL pipeline (v2.0.6) was used to cluster and assemble the Trinity-assembled transcriptome to produce longer and unique contigs (Pertea et al. 2003).

The functional annotation of the assembled unigenes was performed by searching transcripts against the NCBI non-redundant protein Nr, https://www.ncbi.nlm.nih.gov/), Cluster of Orthologous Groups (COG, https://www.ncbi.nlm.nih.gov/research/cog-project/), Swiss-Prot (Boeckmann et al. 2003), and Gene Ontology (GO, http://geneontology.org/) databases. The search was conducted using BLASTx with an E-value cut-off of $1e^{-5}$ (Altschul et al. 1990). The KEGG metabolic pathway database (https://www.kegg.jp/blastkoala/) was searched to assign the molecular and higher-level function to *P. amaryllifolius* transcriptome dataset (Kanehisa et al. 2007). The functional classification of unigenes was performed using the WEGO2.0 (https://wego.genomics.cn/) (Zhang et al. 2018).

SSR locus detection and primer design

The detection of SSRs in unigenes was done using the Perl script MicroSAtellite identification

149 tool (MISA, http://pgrc.ipk-gatersleben.de/misa/misa.html) (Beier et al. 2017). The assembled
150 transcriptome was examined for di-, tri-, tetra, penta-, and hexanucleotide repeat motifs. EST-
151 SSR primers were designed using Primer 3 (http://bioinfo.ut.ee/primer3) (Untergasser et al.
152 2012) according to the following parameters: length 18–23 nucleotides, with 21 optimal; PCR
153 product size 100–400 bp; optimum annealing temperature 55°C; and GC content 40–60%, with
154 50% optimal. For this study, 30 primers were randomly selected for synthesis (Table S2).
155
156 **DNA extraction and genetic variation analysis**
157 This study collected 24 accessions of *P. amaryllifolius* from different locations in Malaysia
158 (Fig. 1; Table S1). The DNA extraction of *P. amaryllifolius* leaves was done using a modified
159 CTAB protocol (Porebski et al. 1997). The quality and quantity of DNA were checked on 1 %
160 gel electrophoresis and a Nanodrop spectrophotometer (Thermo Fisher Scientific, USA),
161 respectively. The 30 EST-SSR primer pairs from the previous section and 10 ISSR primers
162 were validated experimentally through PCR and electrophoresis. A 25 µl final PCR reaction
163 consisted of 1x buffer (Promega, USA), 2.5 mM $MgCl_2$, 0.2 mM dNTP, 1 µM of each primer,
164 0.5 U Taq polymerase (GoTaq, Promega, USA), 25 ng of DNA sample and 15.75 µl of
165 autoclaved distilled water (Kaushal et al. 2017).
166 The PCR for EST-SSR markers was done using a Gradient Thermal cycler (Bio-Rad
167 Laboratories, USA) under the following conditions: initial denaturation of template DNA at
168 95°C for 3 min, followed by 35 cycles of 94°C for 30 s, 57°C – 65°C (depending on the
169 annealing temperature of different primer pairs) for 1 min, 72°C for 1 min, and a final
170 elongation step of 72°C for 10 min. Subsequently, PCR products were separated on 3% (w/v)
171 Metaphor agarose gel (Lonza, Switzerland) electrophoresis.
172 For iPBS markers, sixteen primers with an 18-mer length, which were originally
173 designed by Kalendar et al. (2010), were utilised. These primers were selected due to their

demonstrated high efficiency through PCR (Table S3). First, all primers were amplified with *P. amaryllifolius* accessions to test their efficiency in the amount and quality of bands at 95 °C for 2 min, followed by 34 cycles of 95 °C for 15 s, 50 – 60 °C (Table S6) for 30 s and 72 °C for 60 s, with a final elongation at 72 °C for 5 min. Ultimately, three primers were chosen to analyse 24 *P. amaryllifolius* accessions based on reproducibility and clear visible loci. Each primer was tested twice using genomic DNA from all accessions. The PCR products were separated on 2% agarose gel electrophoresis.

Data analysis

The amplified bands were assessed using manual scoring, where a value of 1 indicated the presence of the band and a value of 0 indicated its absence. In the scoring process, only clearly visible bands that exhibited repetition and were well-distinguished from one another were selected. POPGENE statistical software ver. 1.32 was used in Nei's distance coefficient (h) and Shannon's information index (I) (Yeh 1999). The genetic distance index (D) was used to measure the genetic differences between populations (Nei 1972). The NTSYSpc 2.11 V software package was used to analyse the iPBS-retrotransposon binary matrices (Rohlf 2000). The unweighted pair group method with arithmetic Mean (UPGMA) dendrogram was generated with the SAHN subprogram.

Results

Illumina sequencing and *de novo* assembly of paired-end reads

The sequencing analysis generated 73,687,726 raw reads (7.369 Gb), with the percentage of Q20 bases being 97.18 % (Table 1). The GC content was approximately 41.01 %. After raw read filtering, high-quality reads were assembled into 157,467 unigenes, reaching 170,772,211 bp. All the unigenes were classified based on the length of the sequences, ranging from 301 to

over 3000 bp. The unigenes' average length was approximately 1084.5 bp and N50 of 2,514 (Table 1). Most reads (91.7%) were in the 200-1000 bp range. There were 63,170 unigenes less than 1000 bp in length, 65,339 unigenes ranging from 1000-3000 bp, and 32,363 unigenes with more than 3000 bp (Fig. 2).

Functional annotation and classification of the transcripts

The assembled unigenes were searched against public databases, including NR, SwissProt, COG, TrEMBL, KEGG, and GO, using blastn/x (E-value 10^{-5}) to understand their features and functions. Among these unigenes, 43,891 (27.82%), 76,755 (48.74%), 40,807 (25.91%), 43,991 (27.91%), 6,432 (4.08%) and 22,503 unigenes (14.29%) have significant matches in the NR, SwissProt, COG, TrEMBL, KEGG, and GO databases, respectively. Overall, 112,525 (71.45%) unigenes were annotated in all six databases (Table 2). In the BLASTx search of the NR database, 41,568 unigenes (41.16%) were found to have similarities with sequences from *Aegilops tauschii* (3.69%), *Beta vulgaris* (3.56%), *Cucumis melo* (3.48%), *Glycine max* (3.64%), *Oryza sativa indica* (3.73%), *Oryza sativa japonica* (3.54%), *Physcomitrella patens* (3.03%), *Solanum lypersicum* (3.6%), *Zea mays* (3.67%), and other species (68.09%) (Fig. 3).

In a search against the KEGG database, the unigenes were separated into 5 main categories and were classified into 320 KEGG pathways. Among 5 main categories, metabolism was the biggest category (3,731, 58.01%), followed by genetic information processing (977, 15.19%), organismal systems (660, 10.26%), cellular processes (646, 10.04%) and environmental information processing (660, 10.26%) (Fig. 4).

In this study, *P. amaryllifolius* unigenes were annotated into 3 GO categories and 67 subcategories (Fig. 5). A large number of unigenes were annotated in the cellular component (CC) category and divided into 19 subcategories, of which most were represented by the term of cell, cell part, organelle, organelle part and membrane. In the biological process (BP)

category, the unigenes were clustered into 31 terms at the second level, and the terms, cellular process, metabolic process, and biological regulation, were the top 3 subcategories. In molecular function (MF) category, the unigenes constructed 16 terms at the second level, dominantly represented by binding and catalytic.

Identification of EST-SSRs from *P. amaryllifolius* transcriptome

Approximately 66,820 out of 157,467 unigenes encode 48,816 potential EST-SSRs. Among these unigenes, 13,054 contain multiple SSRs, while 6,596 were identified to be present in compound formation (Table 3). A total of 170,772,211 kilobases (kb) of sequenced data were analysed, revealing a frequency of at least one SSR every 2.6 kb within the expressed part of the *P. amaryllifolius* genome. The SSRs comprised 21,820 mononucleotide (32.65%), 30,428 dinucleotide (45.54%), 11,809 trinucleotide (17.67%), 1,052 quadnucleotide (1.57%), 847 pentanucleotide (1.27%), and 864 hexanucleotide motifs (1.29%) (Fig. S1).

Among the two types of mononucleotide repeat motifs, it was observed that the A/T motif exhibited the highest frequency, accounting for 30.7% of the total number of SSRs (Fig. 6). The frequency of the C/G pattern was found to be less frequent compared to the A/T motif, with a value of 1.95%. The AG/CT dinucleotide repeat motif exhibited the highest prevalence among the other motif types, accounting for 33.94% of the total SSRs. Conversely, the CG/CG motif was found to be the least prevalent, representing only 0.24% of the total SSRs. Among the trinucleotide repeat motifs, the AGG/CCT repeat motif accounted for 4.3% of the total SSRs, making it the most frequently observed pattern. The AAG/CCT repeat motif was the second most common, representing 3.36% of the total SSRs. Following these, the AGC/CTG repeat motif accounted for 1.48% of the total SSRs. The patterns of quad-, penta-, and hexa-nucleotide repeats accounted for 4.13% of the overall SSRs, as seen in Table 4.

SSR marker system

Among 48,186 designed SSR primers, 30 were randomly selected to analyse the polymorphism of 24 *P. amaryllifolius* accessions. Of these, 28 pairs were successfully amplified among these primers, producing the expected band sizes. Two primers (SSR29 and SSR31) produced polymorphic bands from the PCR products (Table 5). The number of polymorphic bands is 2 (SSR19) and 1 (SSR30), with a mean of 1.5 bands per primer. The highest polymorphic information content (PIC) value was 0.14, which was obtained from SSR30. The Nei's genetic diversity for both markers was 0.10 and 0.12, respectively. The average value of Shannon's information index (I) was 0.22.

A UPGMA dendrogram constructed from the two EST-SSR markers generated data has grouped the 24 *P. amaryllifolius* accessions into two clusters (Fig. 7). The first cluster comprised 14 accessions which were collected from Johor, Pulau Pinang, Sarawak and Kedah states. The second cluster contained 10 accessions from Selangor. The principal coordinate analysis (PCoA) presented the existence of two separate groups between the accessions (Fig. 8). The first group consisted of twelve accessions of *P. amaryllifolius* originating from Perak, Johor, Kedah, Sarawak, Sabah, and Pulau Pinang. These accessions were observed to cluster together and positioned in the upper right of the plane. The second group consisted of ten accessions originating from Selangor state and were allocated to the lower left region of the plane.

iPBS marker system

Sixteen iPBS primers were employed in the first screening process to assess polymorphism across the 24 accessions of *P. amaryllifolius*. Three primers (iPBS2399, iPBS2226, and iPBS2238) were chosen for iPBS PCR amplification (Table 6). A total of 10 scorable bands were amplified using these primers, resulting in a polymorphism rate of 73.88% (Table 6). The

primer iPBS2399 exhibited the highest number of band, producing a total of five bands. In contrast, primers iPBS2226 and iPBS2238 yielded three and two bands, respectively. The primer iPBS2399 had the highest PIC value of 0.08, whereas iPBS2238 produced the lowest PIC value of 0.03. The average Nei's genetic diversity and average value of Shannon's information index (I) was 0.03 and 0.10, respectively.

A dendrogram constructed based on the three iPBS markers has grouped the *P. amaryllifolius* accessions into five main clusters (Fig. 9). The first cluster comprised five accessions from Johor. The second cluster contained twelve accessions from Selangor. The third cluster included two accessions of *P. amaryllifolius* from Kedah, the fourth cluster consisted of accessions from Pulau Pinang and the fifth cluster had accessions from Sabah. The PCoA based on iPBS markers revealed the presence of five distinct groups, as depicted in Figure 10. The 24 accessions of *P. amaryllifolius* were distributed into 6 groups. The first group comprised two accessions from Sabah, distributed at the upper left of the plane. Three groups of eight accessions from Sarawak, Kedah and Johor were distributed at the upper right of the plane. The fifth group comprised ten accessions from Selangor and two from Perak, clustered in the lower left plane. The sixth group on the lower left plane had two accessions from Pulau Pinang.

Discussion

De novo assembly and EST-SSR markers development

RNA-seq and *de novo* assembly are powerful tools for gene discovery and generating markers (Unamba et al. 2015). As shown in this study, we successfully generated 73.69 million reads, with an average transcript size of 1395.39 bp, for the analysed *P. amaryllifolius*. In comparison, Bhatt et al. (2021) only reported 29.46 million reads in *P. amaryllifolius*, with a transcript size of 802 bp. Although both assemblies were generated using the Trinity program, different

sequencing depths might cause the discrepancy. However, our assembled *P. amaryllifolius* unigenes (1084.50 bp) are comparable to *Cyamopsis tetragonoloba* (1045.24 bp)(Al-Qurainy et al. 2019) and *Avena sativa L.* (1071.7 bp)(Kim et al., 2022). We achieved a 71.5% annotation rate for the unigenes, probably due to the high amount of unigenes with over 500 base pairs in length within the assembled transcripts. The length of these unigenes appears to affect the annotation rate, with longer sequences increasing the likelihood of identifying BLAST matches in the protein databases (Parchman et al. 2010).

During the BLASTx search in the NR database, we observed low similarity between *P. amaryllifolius* and various plant species, including *Aegilops tauschii, Beta vulgaris, Cucumis melo, Glycine max, Oryza sativa indica, Oryza sativa japonica, Physcomitrella patens, Solanum lypersicum,* and *Zea mays*. This could be attributed to the absence of publicly available whole-genome sequencing of *Pandanus* sp. in existing databases. Our GO analysis showed that the largest groups were the metabolic processes and catalytic activity. The classification of unigenes into 67 subcategories highlighted the diverse transcripts in the *P. amaryllifolius* transcriptome. When analysing the KEGG database, we observed that metabolism was the largest group, followed by genetic information processing, organismal systems, cellular processes and environmental information processing (Figure 6). These findings revealed the possibility of generating high-quality transcriptome data for non-model plants with large genomes using high-throughput technology.

DNA markers derived from transcriptomic sequences often outperform those derived from genomic sequence data (Taheri et al. 2019). In this study, we successfully identified EST-SSR markers (30.06%), with a distribution density of one SSR per 2.6 kb, which was remarkably higher than *Curcuma alismatifolia* (12.57%) (Taheri et al. 2019), *Panicum sumatrense* (11.25%) (Desai et al. 2021), *Mucuna pruriens* (9.3%) (Sathyanarayana et al. 2017), *Magnolia wufengensis* (8.7%) (Wang et al. 2019), *Coriandrum sativum* L. (21.91%)

(Tulsani et al. 2020), and *Polygonatum sibiricum* (25.86%) (Yang et al. 2022), but lower than *Glycyrrhiza uralensis* (60.10%) (Liu et al. 2015) and *Gentiana macrophylla* (30.73%) (Wang et al. 2015). The frequency and abundance variations of SSRs between different species can be attributed to several factors, including the criteria utilised for SSR detection, the size of the dataset, the techniques employed for database mining, the specific species under investigation, and the scale of the unigene assembly dataset (Varshney et al. 2005).

In this study, we identified six distinct repeat motifs. We found that dinucleotide repeats (45.54%) were the most prevalent type, followed by mononucleotide repeats (32.64%) and trinucleotide (17.67%). In contrast, hexa- (1.29%), penta- (1.29%), and quad nucleotide repeats (1.57%) were rare. The present analysis revealing a substantial prevalence of dinucleotide repeats aligns with other research findings on *Mentha piperita* (Kumar et al. 2015)*, Camellia sinensis* (Liu et al. 2018), *Jatropha curcas* L. (Kumar Yadav et al. 2011), and *Capsicum frutescens* (Zhong et al. 2021). According to Kumpatla and Mukhopadhyay (Kumpatla and Mukhopadhyay 2005), the abundance of dinucleotide repeat sequences may be due to the UTRs' overexpression compared to open reading frames. The AG/CT motif (22,697, 33.94%) was the most common of the dinucleotide repeats, followed by the AT/AT motif (4,649, 6.96%) (Table 4). These findings were similar to previous studies on *Polygonatum sibiricum* (Yang et al. 2022) and *Glycyrrhiza uralensis* (Liu et al. 2015). The dominant motifs for mononucleotide were A/T (30.70%), followed by C/G (1.95%). The prevalent of A/T was also reported in *Glycyrrhiza uralensis* (Liu et al. 2018), *Punica granatum L* (Hasnaoui et al. 2012), and *Eucommia ulmoides* (Zhang et al. 2016). According to Gur-Arie et al. (2000), a high number of A/T bases implies a potential correlation between the phenomena and the ease of DNA melting in repeat sequences that are abundant in A/T bases.

Genetic diversity analysis using EST-SSR and iPBS

Genetic diversity in plant species can be effectively assessed using specific molecular markers, depending on the research objective and the genetic makeup of the plant species. In this study, both the EST-SSR and iPBS marker systems effectively differentiate *P. amaryllifolius* accessions into separate groups. Nevertheless, the markers employed in the study revealed a low level of polymorphism, as indicated by the PIC value and the previous study by Wakte et al. (2012). PIC is a marker index to estimate the discriminating ability and informativeness of markers employed in the analysis. It allows for measuring the effectiveness of polymorphic loci in detecting genetic diversity among different plant varieties (Botstein et al. 1980). The PIC values are often categorised as low ($PIC < 0.25$), medium ($0.5 > PIC > 0.25$), or high ($PIC > 0.5$) (Botstein et al. 1980). They have been widely used to assess the informativeness EST-SSR and iPBS markers in different cultivated species (Eren et al. 2022; Guo et al. 2014; Sharma et al. 2020; Shimira et al. 2021). For example, several studies reported low PIC value for various plant species (Singh et al. 2013; Ates et al. 2019; Yaldiz et al. 2018; Shirmohammadli et al. 2018).

The present study observed a lower average number of polymorphic EST-SSR and iPBS-retrotransposon bands compared to those reported by Kizilgeci et al. (2022), Borna et al. (2017), Aydın et al. (2020) and Karagoz et al. (2020). This could be because EST-SSR and iPBS-retrotransposon primers for *P. amaryllifolius* genome are more conserved than other studied plants. To further assess the genetic diversity among the varieties and evaluate the performance of the examined markers, the Shannon's information index (I) and genetic diversity (h) parameters were calculated. The results revealed that these loci are informative, and EST-SSR and iPBS markers efficiently determine the genetic diversity and relationships of *P. amaryllifolius*.

The UPGMA dendrogram and PCoA produced from the DNA profile displayed the genetic diversity of the *P. amaryllifolius* accessions. The comparison of population clustering

between EST-SSR and iPBS markers was not similar. This study showed that the iPBS marker system is more effective than EST-SSR markers in distinguishing genetic diversity in *P. amaryllifolius*. One possible reason for this could be the location of the markers in the genome. For instance, iPBS markers are generated from genomic regions (Kalendar et al. 2010), while EST-SSR markers are derived from expressed regions of the genome (Parthiban et al. 2018). For this reason, iPBS markers can potentially capture a wider range of genetic variation than EST-SSR markers, which are limited to variations in expressed regions. Based on the iPBS marker system, the accessions of *P. amaryllifolius* exhibited minimal genetic variation, although they displayed a tendency to cluster according to their respective geographical origins. This finding indicates that accessions from the same geographic region exhibit no genetic variation, but slight genetic diversity may be detected among accessions from different states. Accessions from Pulau Pinang and Sabah might be distributed from different sources since they were clustered differently from other accessions. *P. amaryllifolius* is said to be the original native of Moluccas (The Maluku Islands) since the only flowering specimen was found in that geographical region (Stone 1987).

Conclusions

The *de novo* assembly of the *P. amaryllifolius* transcriptome, facilitated by RNA-seq technology, produced a substantial number of unigenes. These unigenes were used to develop suitable EST-SSR primers. Based on the genetic diversity analysis, we found that the iPBS marker system showed a higher efficacy in evaluating polymorphism in *P. amaryllifolius* compared to the EST-SSR marker system. Furthermore, the availability of the transcriptome dataset would provide a comprehensive collection of distinct genes, together with profound understanding of their respective pathways. These findings could be a significant resource for genetic investigations on *P. amaryllifolius.* However, considering that genetic analysis was

399 conducted on only 24 individuals from different regions of Malaysia, it is essential to conduct
400 further studies with a larger sample size to analyse the genetic diversity and structure of the
401 species.
402
403 **References**

404 Al-Qurainy F, Alshameri A, Gaafar AR, Khan S, Nadeem M, Alameri AA, Tarroum M, Ashraf
405 (2019) Comprehensive stress-based de novo transcriptome assembly and annotation of guar
406 (*Cyamopsis tetragonoloba* (L.) Taub.): an important industrial and forage crop. Int J Genomics
407 2019. https://doi.org/10.1155/2019/7295859
408
409 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search
410 tool. J Mol Biol 215:403-410. https://doi.org/10.1016/S0022-2836(05)80360-2
411
412 Amnan MAM, Aizat WM, Khaidizar FD, Tan BC (2022) Drought stress induces morpho
413 physiological and proteome changes of *Pandanus amaryllifolius*. Plants 11:221.
414 https://doi.org/10.3390/plants11020221
415
416 Amom T, Nongdam P (2017) The use of molecular marker methods in plants: a review. Int J
417 Curr Res Rev 9:1-7. https://doi.org/10.7324/IJCRR.2017.9171
418
419 Anderson JA, Churchill G, Autrique J, Tanksley S, Sorrells M (1993) Optimizing
420 parental selection for genetic linkage maps. Genome 36:181-186.
421 https://doi.org/10.1139/g93-024

422

Ates D, Altinok HH, Ozkuru E, Ferik F, Erdogmus S, Can C, Tanyolac MB (2019) Population structure and linkage disequilibrium in a large collection of *Fusarium oxysporum* strains analysed through iPBS markers. J Phytopathol 167:576-590. https://doi.org/10.1111/jph.12848

Aydın F, Özer G, Alkan M, Çakır İ (2020) The utility of iPBS retrotransposons markers to analyze genetic variation in yeast. Int J Food Microbiol 325:108647. https://doi.org/10.1016/j.ijfoodmicro.2020.108647

Bayat M, Amirnia R, Özkan H, Gedik A, Ates D, Tanyulac B, Rahimi M (2018) Diversity and phylogeny of saffron (*Crocus sativus* L.) accessions based on IPBS markers. Genetika 50:33-44. https://doi.org/10.2298/GENSR1801033B

Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) MISA-web: a web server for microsatellite prediction. Bioinform 33:2583-2585. https://doi.org/10.1093/bioinformatics/btx198

Bhatt V, Barvkar VT, Furtado A, Henry RJ, Nadaf A (2021) Fragrance in *Pandanus amaryllifolius* Roxb. despite the presence of a betaine aldehyde dehydrogenase 2. Int J Mol Sci 22:6968. https://doi.org/10.3390/ijms22136968

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucl Acids Res 31:365-370. https://doi.org/10.1093/nar/gkg095

448 Bonchev G, Dušinský R, Hauptvogel P, Gaplovská-Kyselá K, Švec M (2019) On the diversity
449 and origin of the barley complex agriocrithon inferred by iPBS transposon markers. Genet
450 Resour Crop Evol 66:1573-1586.
451
452 Boopathi NM (2020) Genotyping of mapping population. In Genetic mapping and marker
453 assisted selection. Springer. https://doi.org/10.1007/978-81-322-0958-4_8
454
455 Borna F, Luo S, Ahmad N, Nazeri V, Shokrpour M, Trethowan R (2017) Genetic diversity in
456 populations of the medicinal plant Leonurus cardiaca L. revealed by inter-primer binding site
457 (iPBS) markers. Genet Resour Crop Evol 64:479-492. https://doi.org/10.1007/s10722-016-
458 0373-4
459
460 Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map
461 in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314.
462
463 Cai K, Zhu L, Zhang K, Li L, Zhao Z, Zeng W, Lin X (2019) Development and characterization
464 of EST-SSR markers from RNA-Seq data in *Phyllostachys violascens*. Front Plant Sci 10.
465 https://doi.org/10.3389/fpls.2019.00050
466
467 Chiabchalard A, Nooron N (2015) Antihyperglycemic effects of *Pandanus amaryllifolius*
468 Roxb. leaf extract. Pharmacogn Mag 11:117-22. https://doi.org/10.4103/0973-1296.149724
469
470 Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM (2010) The Sanger FASTQ file format for
471 sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res
472 38:1767-177. https://doi.org/10.1093/nar/gkp1137

Coutinho JP, Carvalho A, Martín, A, Lima-Brito J (2018) Molecular characterization of Fagaceae species using inter-primer binding site (iPBS) markers. Mol Biol Rep 45:133-142. https://doi.org/10.1007/s11033-018-4146-3

Desai H, Hamid R, Ghorbanzadeh Z, Bhut N, Padhiyar SM, Kheni J, Tomar RS (2021) Genic microsatellite marker characterization and development in *Panicum sumatrense* using transcriptome sequencing. Sci Rep 11:1-14. https://doi.org/10.1038/s41598-021-98236-7

Eren B, Keskin B, Demirel F, Demirel S, Türkoğlu A, Yilmaz A, Haliloğlu K (2022) Assessment of genetic diversity and population structure in local alfalfa genotypes using iPBS molecular markers. Genet Resour Crop Evol 1-12. https://doi.org/10.1007/s10722-022-01425-6

Gao J, Li WB, Liu HF, Chen FB (2019) De novo transcriptome sequencing of Raphanus sativus L fleshy roots: analysis of major genes involved in the anthocyanin synthesis pathway. BMC Mol Cell Biol 20:1-12. https://doi.org/10.1186/s12860-019-0205-0

Ghasemzadeh A and Jaafar HZ (2013) Profiling of phenolic compounds and their antioxidant and anticancer activities in pandan (*Pandanus amaryllifolius* Roxb.) extracts from different locations of Malaysia. BMC complementary and alternative medicine 13(1-9).

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q (2011) Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol 29:644. https://doi.org/10.1038/nbt.1883

Guo DL, Guo MX, Hou XG, Zhang GH (2014) Molecular diversity analysis of grape varieties based on iPBS markers. Biochem Syst Ecol 52:27-32. https://doi.org/10.1016/j.bse.2013.11.017

Gur-Arie R, Cohen CJ, Eitan Y, Shelef L, Hallerman EM, Kashi Y (2000) Simple sequence repeats in Escherichia coli: abundance, distribution, composition, and polymorphism. Genome Res 10:62-71. https://doi.org/10.1101/gr.10.1.62

Hasnaoui N, Buonamici A, Sebastiani F, Mars M, Zhang D, Vendramin GG (2012) Molecular genetic diversity of *Punica granatum* L (pomegranate) as revealed by microsatellite DNA markers (SSR). Gene 493:105-112. https://doi.org/10.1016/j.gene.2011.11.035

Kaban N, Putri PS (2020) Pemberian air daun pandan terhadap penurunan kadar gula darah pada pasien diabetes. JKM (Jurnal Kebidanan Malahayati) 6:493-496.

Kalendar R, Antonius K, Smýkal P, Schulman AH (2010) iPBS: a universal method for DNA fingerprinting and retrotransposon isolation. Theor Appl Genet 121:1419-1430. https://doi.org/10.1007/s00122-010-1419-x

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T (2007) KEGG for linking genomes to life and the environment. Nucleic Acids Res 36:D480-D484. https://doi.org/10.1093/nar/gkm882

522 Karagoz H, Cakmakci R, Hosseinpour A, Ozkan G, Haliloglu K (2020) Analysis of genetic
523 variation and population structure among of oregano (*Origanum acutidens* L.) accessions
524 revealed by agro-morphological traits, oil constituents, and retrotransposon-based inter-primer
525 binding sites (iPBS) markers. Genet Resour Crop Evol 67:1367-1384.
526 https://doi.org/10.1007/s10722-020-01016-5
527
528 Kaushal A, Singh A, Jeena AS (2017) Genetic diversity in tomato (*Solanum lycopersicum* L)
529 genotypes revealed by simple sequence repeats (SSR) markers. J Appl Nat Sci 9:966-973.
530
531 Kim TH, Yoon YM, Park JC, Park JH, Kim KH, Kim YK, Son JH, Park TI (2022) De Novo
532 Transcriptome Assembly and SNP Discovery for the Development of dCAPS Markers in Oat.
533 Agronomy 12:184. https://doi.org/10.3390/agronomy12010184
534
535 Kizilgeci F, Bayhan B, Türkoğlu A, Haliloglu K, Yildirim M (2022) Exploring genetic
536 diversity and population structure of five *Aegilops* species with inter-primer binding site (iPBS)
537 markers. Mol Biol Rep 49:8567-8574. https://doi.org/10.1007/s11033-022-07545-x
538
539 Kong Q, Liu Y, Xie J, Bie Z (2020) Development of simple sequence repeat markers from de
540 novo assembled transcriptomes of pumpkins. Plant Mol Biol Rep 38:130-136.
541 https://doi.org/10.1007/s11105-020-01282-3
542
543 Kumar B, Kumar U, Yadav HK (2015) Identification of EST–SSRs and molecular diversity
544 analysis in *Mentha piperita*. Crop J 3:335-342. https://doi.org/10.1016/j.cj.2015.05.008
545

546 Kumar Yadav H, Ranjan A, Asif MH, Mantri S, Sawant SV, Tuli R (2011) EST-derived SSR
547 markers in *Jatropha curcas* L.: development, characterization, polymorphism, and
548 transferability across the species/genera. Tree Genet Genomes 7:207-219.
549 https://doi.org/10.1007/s11295-010-0349-2
550
551 Kumpatla SP, Mukhopadhyay S (2005) Mining and survey of simple sequence repeats in
552 expressed sequence tags of dicotyledonous species. Genome 48:985-998.
553 https://doi.org/10.1139/g05-052
554
555 Laluces HMC, Nakayama A, Nonato MG, dela Cruz TE, Tan MA (2015) Antimicrobial
556 alkaloids from the leaves of *Pandanus amaryllifolius*. J Appl Pharm Sci 5:151-153.
557
558 Lee BL, Su J, Ong CN (2004) Monomeric C18 chromatographic method for the liquid
559 chromatographic determination of lipophilic antioxidants in plants. J Chromatogr A 1048:263-
560 267. https://doi.org/10.1016/j.chroma.2004.08.097
561
562 Liu S, An Y, Li F, Li S, Liu L, Zhou Q, Zhao S, Wei C (2018) Genome-wide identification of
563 simple sequence repeats and development of polymorphic SSR markers for genetic studies in
564 tea plant (*Camellia sinensis*). Mol Breeding 38:1-13. https://doi.org/10.1007/s11032-017-
565 0742-3
566
567 Liu Y, Zhang P, Song M, Hou J, Qing M, Wang W, Liu C (2015) Transcriptome Analysis and
568 Development of SSR Molecular Markers in *Glycyrrhiza uralensis* Fisch. PLoS One
569 10:e0143017. https://doi.org/10.1371/journal.pone.0143017
570

571 Mathew KM, Reshma R, Geethu M, Rithin V, Sabu K, Nadiya F, Noushad MA, Dharan SS,
572 Rao Y, Remashree A (2019) Data on small cardamom transcriptome associated with capsule
573 rot disease. Data Brief 27. https://doi.org/10.1016/j.dib.2019.104654
574
575 Mohd Amnan MA, Teo WFA, Aizat WM, Khaidizar FD, Tan BC (2023) Foliar application of
576 oil palm wood vinegar enhances *Pandanus amaryllifolius* tolerance under drought
577 stress. Plants 12:785. https://doi.org/10.3390/plants12040785
578
579 Nei M (1972) Genetic distance between populations. Am Nat 106:283-292.
580 https://doi.org/10.1086/282771
581
582 Parchman TL, Geist KS, Grahnen JA, Benkman CW, Buerkle CA (2010) Transcriptome
583 sequencing in an ecologically important tree species: assembly, annotation, and marker
584 discovery. BMC Genomics 11:1-16. https://doi.org/10.1186/1471-2164-11-180
585
586 Parthiban S, Govindaraj P, Senthilkumar S (2018) Comparison of relative efficiency of
587 genomic SSR and EST-SSR markers in estimating genetic diversity in sugarcane. 3 Biotech
588 8:1-12. https://doi.org/10.1007/s13205-017-1092-4
589
590 Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung
591 F, Parvizi B (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast
592 clustering of large EST datasets. Bioinform 19:651-652.
593 https://doi.org/10.1093/bioinformatics/btg034
594

595 Porebski S, Bailey LG, Baum BR (1997) Modification of a CTAB DNA extraction protocol
596 for plants containing high polysaccharide and polyphenol components. Plant Mol Biol Rep
597 15:8-15. https://doi.org/10.1007/BF02772108
598
599 Reshidan NH, Abd Muid S, Mamikutty N (2019) The effects of *Pandanus amaryllifolius*
600 (Roxb.) leaf water extracts on fructose-induced metabolic syndrome rat model. BMC
601 Complement Altern Med 19:1-13. https://doi.org/10.1186/s12906-019-2436-9
602
603 Rohlf F (2000) NTSYS-pc: Numerical taxonomy and multivariate analysis system version 2.2.
604 Exeter Publications. Setauket, NY.
605
606 Sathyanarayana N, Pittala RK, Tripathi PK, Chopra R, Singh HR, Belamkar V, Bhardwaj PK,
607 Doyle JJ, Egan AN (2017) Transcriptomic resources for the medicinal legume *Mucuna*
608 *pruriens*: de novo transcriptome assembly, annotation, identification and validation of EST-
609 SSR markers. BMC Genomics 18:1-18. https://doi.org/10.1186/s12864-017-3605-0
610
611 Sharma H, Kumar P, Singh A, Aggarwal K, Roy J, Sharma V, Rawat S (2020) Development
612 of polymorphic EST-SSR markers and their applicability in genetic diversity evaluation in
613 *Rhododendron arboreum*. Mol Biol Rep 47:2447-2457. https://doi.org/10.1007/s11033-020-
614 05580-5
615
616 Shimira F, Boyaci HF, Çilesiz Y, Nadeem MA, Baloch FS, Taşkin H (2021) Exploring the
617 genetic diversity and population structure of scarlet eggplant germplasm from Rwanda through
618 iPBS-retrotransposon markers. Mol Biol Rep 48:6323-6333. https://doi.org/10.1007/s11033-
619 021-06350-x

620

621 Shirmohammadli S, Sabouri H, Ahangar L, Ebadi AA, Sajjadi SJ (2018) Genetic diversity and
622 association analysis of rice genotypes for grain physical quality using iPBS, IRAP, and ISSR
623 markers. J Genet Res 15:122-129. https://10.22080/JGR.2019.15415.1115

624

625 Simms D, Cizdziel PE, Chomczynski P (1993) TRIzol: A new reagent for optimal single-step
626 isolation of RNA. Focus 15:532-535.

627

628 Singh N, Choudhury DR, Singh AK, Kumar S, Srinivasan K, Tyagi R, Singh N, Singh R (2013)
629 Comparison of SSR and SNP markers in estimation of genetic diversity and population
630 structure of Indian rice varieties. PloS One 8:e84136.
631 https://doi.org/10.1371/journal.pone.0084136

632

633 Stone BC (1978) Studies in Malesian Pandanaceae XVII on the taxonomy of '*Pandan Wangi*'A
634 Pandanus cultivar with scented leaves. Econ Bot 32:285-293.

635

636 Taheri S, Abdullah TL, Rafii M, Harikrishna JA, Werbrouck SP, Teo CH, Sahebi M, Azizi P
637 (2019) De novo assembly of transcriptomes, mining, and development of novel EST-SSR
638 markers in *Curcuma alismatifolia* (Zingiberaceae family) through Illumina sequencing. Sci
639 Rep 9:1-14. https://doi.org/10.1038/s41598-019-43958-6

640

641 Tattersall EA, Ergul A, AlKayal F, DeLuc L, Cushman JC, Cramer GR (2005) Comparison of
642 methods for isolating high-quality RNA from leaves of grapevine. Am J Enol Vitic 56:400-
643 406. https://doi.org/10.5344/ajev.2005.56.4.400

644

Tulsani NJ, Hamid R, Jacob F, Umretiya NG, Nandha AK, Tomar RS, Golakiya BA (2020) Transcriptome landscaping for gene mining and SSR marker development in Coriander (*Coriandrum sativum* L.). Genomics 112:1545-1553. https://doi.org/10.1016/j.ygeno.2019.09.004

Unamba CI, Nag A, Sharma RK (2015) Next generation sequencing technologies: the doorway to the unexplored genomics of non-model plants. Front Plant Sci 6:1074. https://doi.org/10.3389/fpls.2015.01074

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3 - new capabilities and interfaces. Nucleic Acids Res 40:e115. https://doi.org/10.1093/nar/gks596

Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. Trends Biotechnol 23:48-55. https://doi.org/10.1016/j.tibtech.2004.11.005

Wakte KV, Nadaf AB, Krishnan S, Thengane RJ (2007) Studies on lower epidermal papillae, the site of storage of basmati rice aroma compounds in *Pandanus amaryllifolius* Roxb. Curr Sci 238-242.

Wakte KV, Zanan RL, Saini A, Jawali N, Thengane RJ, Nadaf AB (2012) Genetic diversity assessment in *Pandanus amaryllifolius* Roxb. populations of India. Genet Res Crop Evol 59:1583-1595. https://doi.org/10.1007/s10722-012-9882-y

Wang L, Gong X, Jin L, Li H, Lu J, Duan J, Ma L (2019) Development and validation of EST-SSR markers of *Magnolia wufengensis* using de novo transcriptome sequencing. Trees 33:1213-1223. https://doi.org/10.1007/s00468-019-01853-2

Wang L, Wang Z, Chen J, Liu C, Zhu W, Wang L, Meng L (2015) De novo transcriptome assembly and development of novel microsatellite markers for the traditional Chinese medicinal herb, *Veratrilla baillonii* Franch (Gentianaceae). Evol Bioinformatics 11:EBO.S20942. https://doi.org/10.4137/EBO.S20942

Wang X, Yang S, Chen Y, Zhang S, Zhao Q, Li M, Gao Y, Yang L, Bennetzen JL (2018) Comparative genome-wide characterization leading to simple sequence repeat marker development for Nicotiana. BMC Genomics 19:1-12. https://doi.org/10.1186/s12864-018-4878-4

Wongpornchai S (2006) Pandan wangi. In Handbook of herbs and spices 453-459, Elsevier.

Yahya F, Fryer PJ, Bakalis S (2011) The absorption of 2-acetyl-1-pyrroline during cooking of rice (*Oryza sativa* L.) with Pandan (*Pandanus amaryllifolius* Roxb.) leaves. Procedia Food Sci 1:722-728. https://doi.org/10.1016/j.profoo.2011.09.109

Yaldiz G, Camlica M, Nadeem MA, Nawaz MA, BALOCH FS (2018) Genetic diversity assessment in Nicotiana tabacum L. with iPBS-retrotransposons. Turk J Agric For 42:154-164.

Yang Q, Jiang Y, Wang Y, Han R, Liang Z, He Q, Jia Q (2022) SSR Loci Analysis in Transcriptome and Molecular Marker Development in *Polygonatum sibiricum*. Biomed Res Int 2022. https://doi.org/10.3906/tar-1708-32

Yeh FC (1999) POPGENE (version 1.3.1). Microsoft Window-Based Freeware for Population Genetic Analysis. http://www.ualberta.ca/~fyeh/.

Zhang H, He L, Cai L (2018) Transcriptome sequencing: RNA-seq. In Computational Systems Biol 15-27, Springer. https://doi.org/10.1007/978-1-4939-7717-8_2

Zhang W, Li Y, Zhao J, Wu C, Ye S, Yuan W (2016) Isolation and characterization of microsatellite markers for *Eucommia ulmoides* (Eucommiaceae), an endangered tree, using next-generation sequencing. Genet Mol Res 15. https://doi.org/10.4238/gmr.15027789

Zhang Z, Xie W, Zhao Y, Zhang J, Wang N, Ntakirutimana F, Yan J, Wang Y (2019) EST-SSR marker development based on RNA-sequencing of *E. sibiricus* and its application for phylogenetic relationships analysis of seventeen *Elymus* species. BMC Plant Biol 19:1-18. http://doi.org/10.1186/s12870-019-1825-8

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 7:187. https://doi.org/10.3390/horticulturae7070187

720

721

722

723

724 **List of figure captions**

725 **Fig. 1.** Geographical location of P. amaryllifolius samples collected in Malaysia.

726 **Fig. 2.** Length distribution of the assembled unigenes in P. amaryllifolius.

727 **Fig. 3.** Species-based distribution of the top BLASTx hits for each assembled unigenes.

728 **Fig. 4.** KEGG enrichment for the transcriptome of Pandanus amaryllifolius. The groups are

729 labelled as follows: Metabolism categories, Genetic information processing categories,

730 Environmental information processing categories, Cellular process categories, and

731 Organismal systems categories.

732 **Fig. 5.** Gene ontology (GO) classification of assembled unigenes of P. amaryllifolius.

733 Results are summarized in three main categories: molecular function, cellular component,

734 biological process. The x-axis indicates the subgroups in GO annotation while the y-axis

735 indicates the percentage of specific categories of genes in each main category.

736 **Fig. 6.** The frequency distribution and repeat numbers of SSRs based on motif sequence type.

737 **Fig. 7.** UPGMA dendrogram of 24 P. amaryllifolius accessions based on two EST-SSR
738 markers.
739
740 **Fig. 8.** Principal coordinate analysis of 24 *P. amaryllifolius* accessions based on two EST-
741 SSR markers.
742
743 **Fig. 9.** UPGMA dendrogram of 24 P. amaryllifolius accessions based on three iPBS markers.
744
745 **Fig. 10.** Principal coordinate analysis of 24 *P. amaryllifolius* accessions based on three iPBS
746 markers.
747
748 **Fig. S1.** Distribution of different SSR repeat motif types.
749
750
751

752

List of table captions

Table 1. Summary of assembly results for *Pandanus amaryllifolius.*

Table 2. Summary of functional annotation of unignenes *P. amaryllifolius.*

Table 3. Summary of EST-SSRs identified from *P. amaryllifolius.*

Table 4. Summary of EST-SSRs identified from the unigenes of *P. amaryllifolius.*

Table 5. Characteristics of the two polymorphic EST-SSR markers in 24 *P. amaryllifolius* samples.

Table 6. Characteristics of the three polymorphic iPBS markers in 24 *P. amaryllifolius* samples.

Table S1. List of samples of *P. amaryllifolius* collected from various geographical locations in Malaysia.

Table S2. Characteristics of EST-SSR primers designed for analysing genetic diversity of *P. amaryllifolius.*

Table S3. Characteristics of iPBS primers selected for analysing genetic diversity of *P. amaryllifolius*.

Supplementary File caption

Supplementary File 1: EST-SSR markers generated from *P. amaryllifolius* transcriptome.

Figures

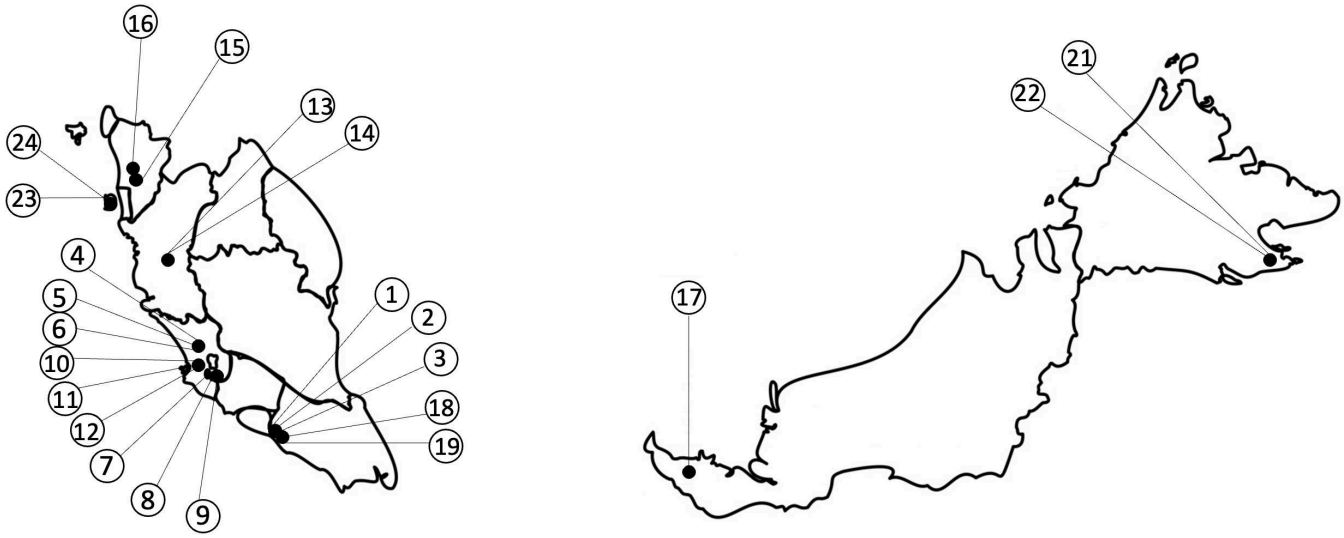


Figure 1

Geographical locationof P. amaryllifolius samples collected in Malaysia.

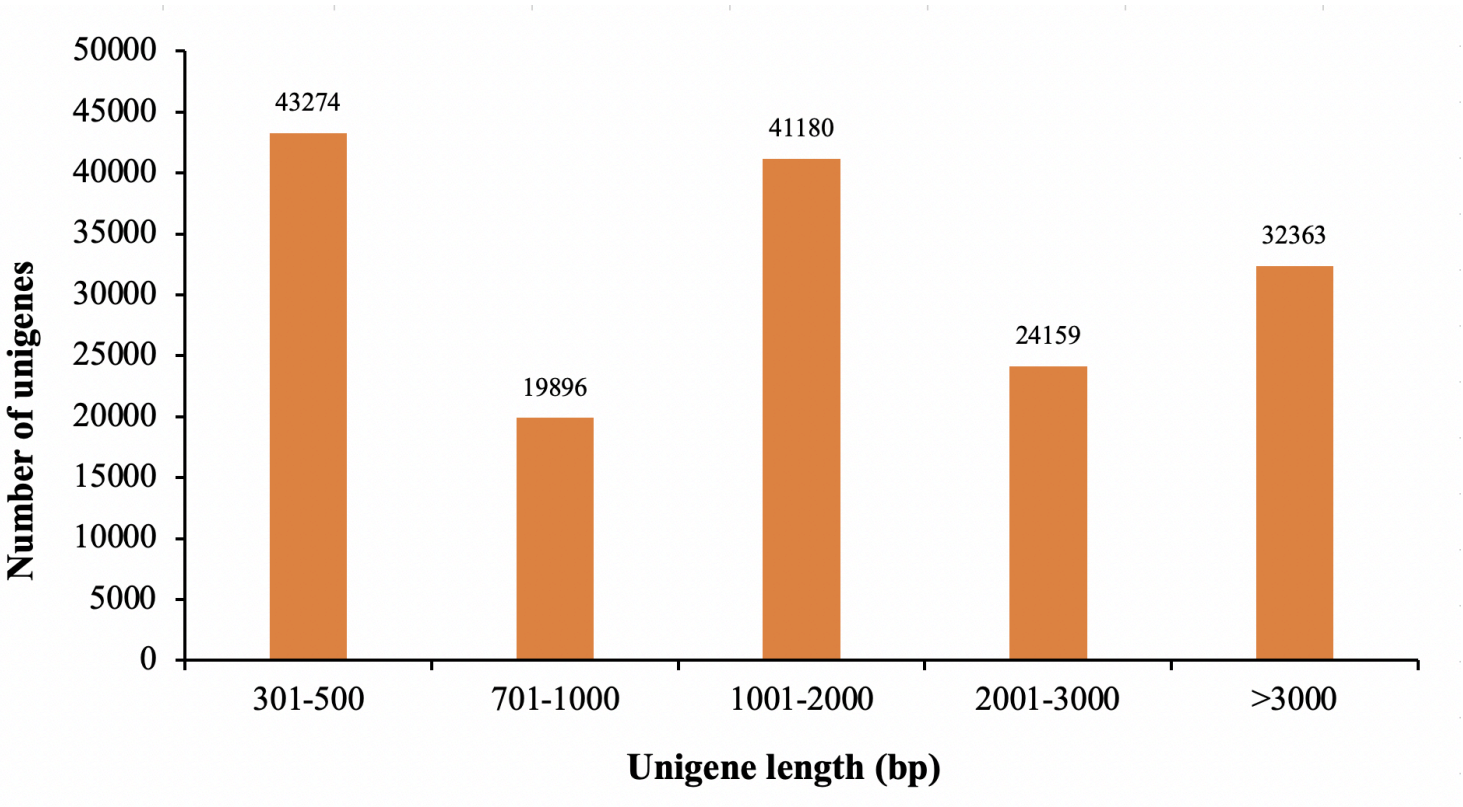


Figure 2

Length distribution of the assembled unigenes in P. amaryllifolius.

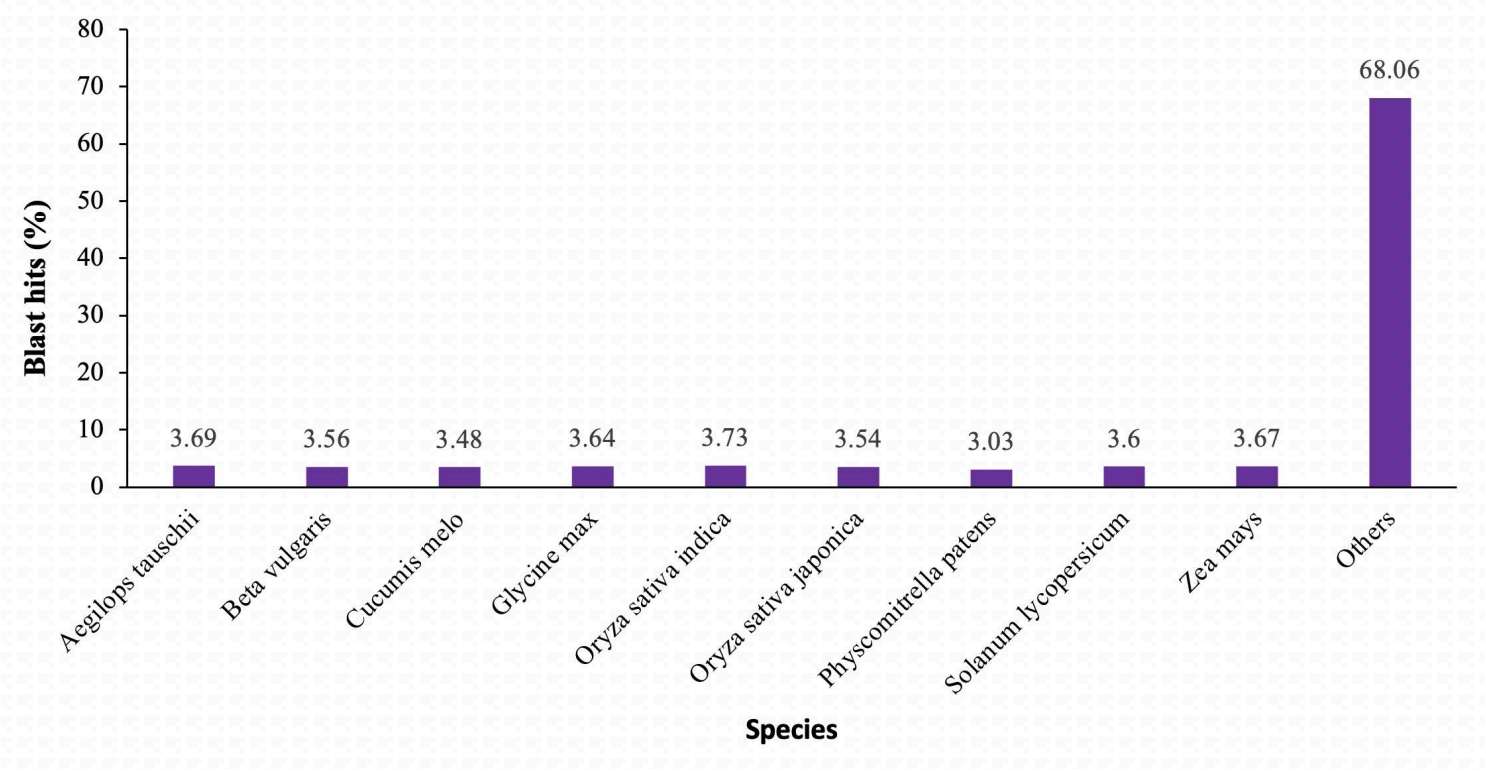


Figure 3

Species-based distribution of the top BLASTx hits for each assembled unigenes.

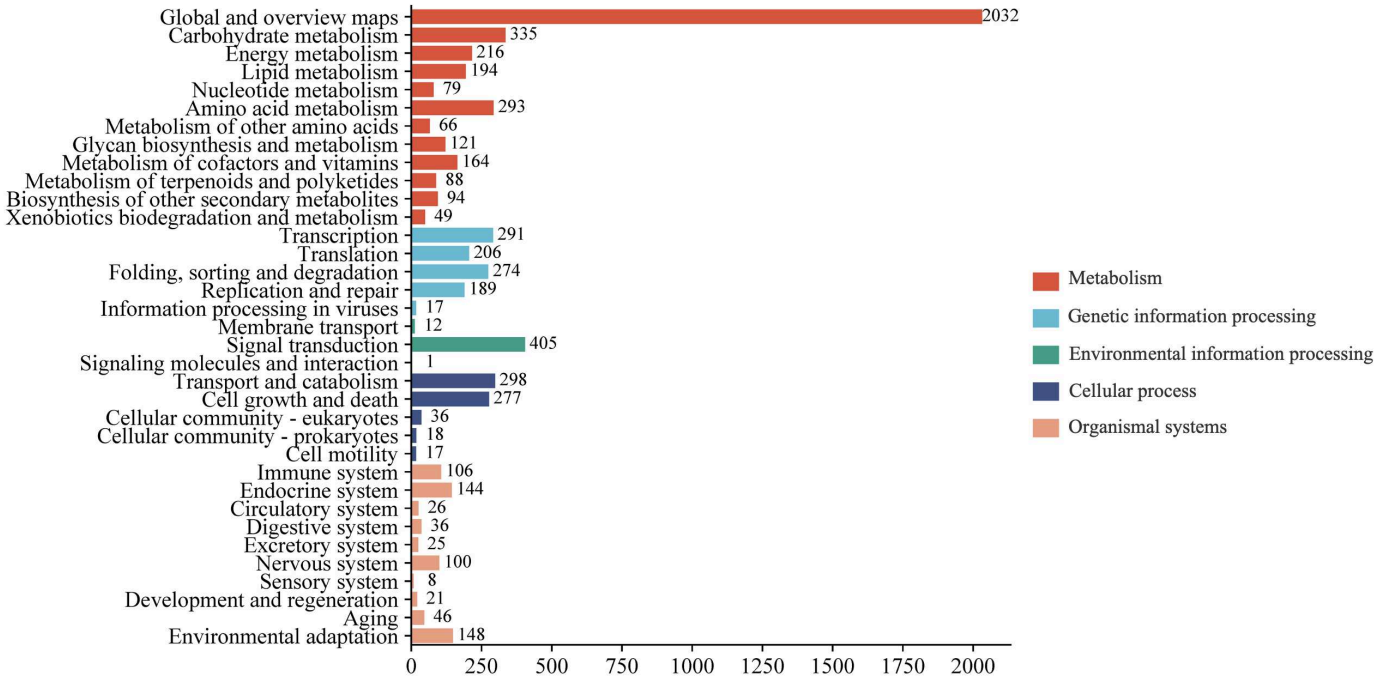


Figure 4

KEGG enrichment for the transcriptome of Pandanus amaryllifolius. The groups are labelled as follows: Metabolism categories, Genetic information processing categories, Environmental information processing categories, Cellular process categories, and Organismal systems categories.

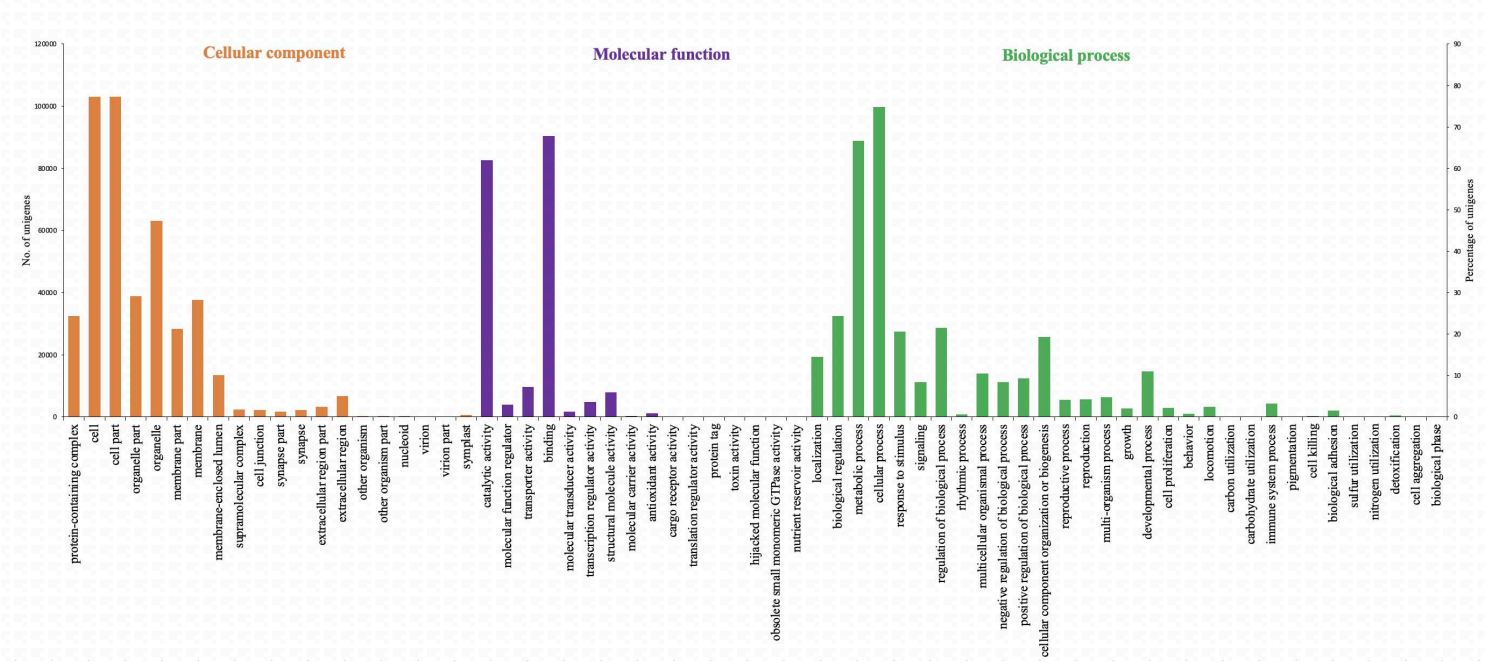


Figure 5

Gene ontology (GO) classification of assembled unigenes of P. amaryllifolius. Results are summarized in three main categories: molecular function, cellular component, biological process. The x-axis indicates the subgroups in GO annotation while the y-axis indicates the percentage of specific categories of genes in each main category.

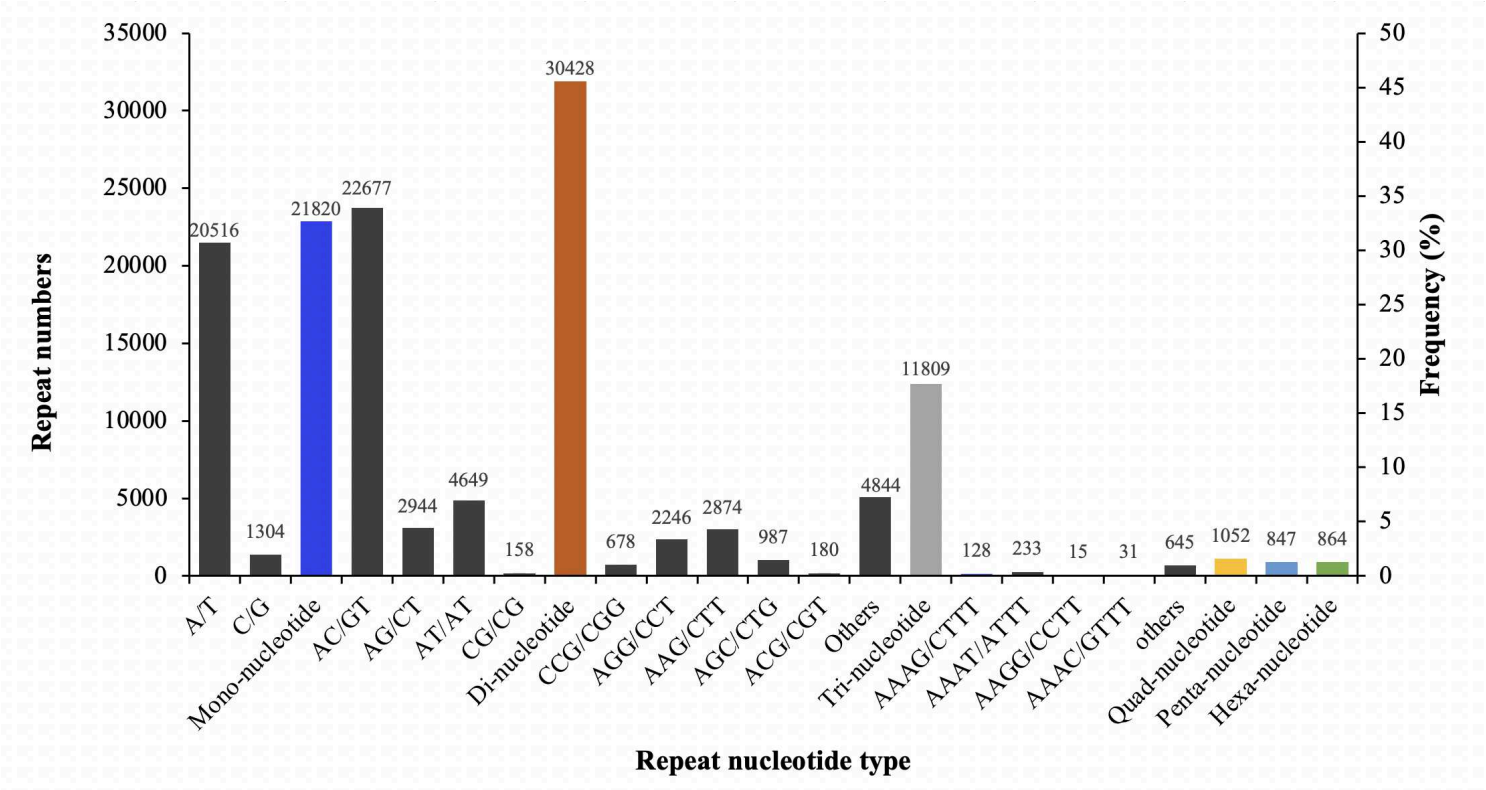


Figure 6

The frequency distribution and repeat numbers of SSRs based on motif sequence type.

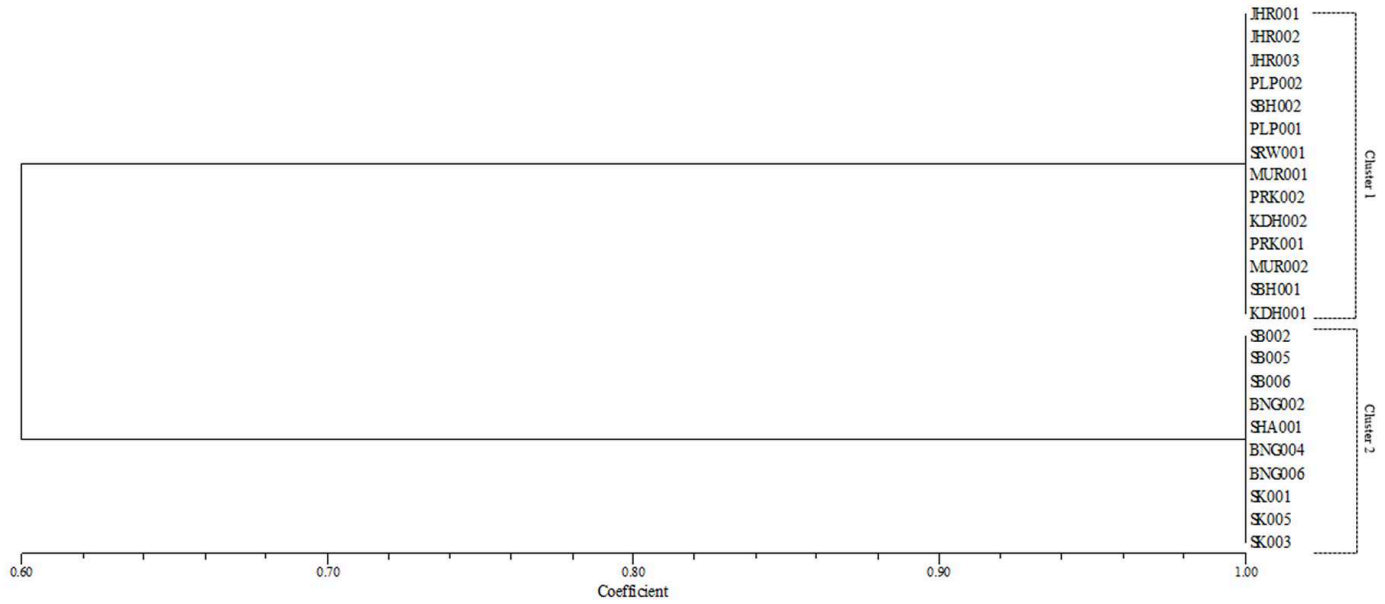


Figure 7

UPGMA dendrogram of 24 P. amaryllifolius accessions based on two EST-SSR markers.

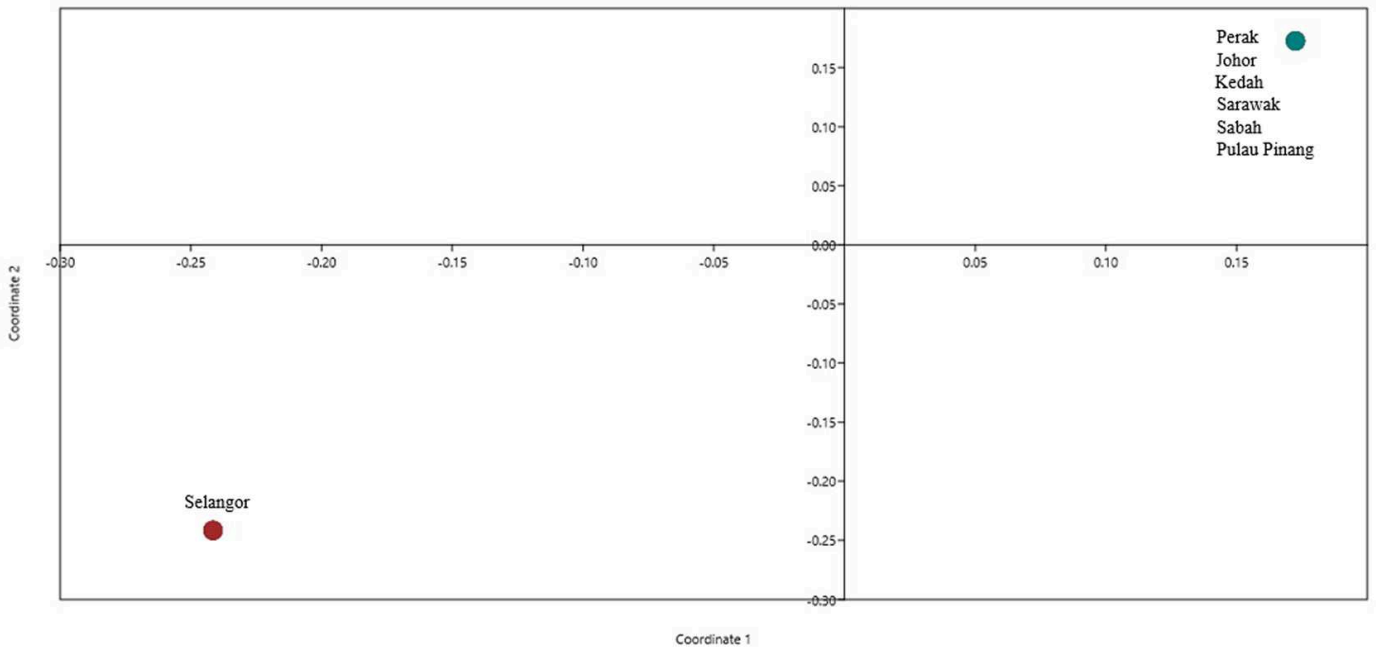


Figure 8

Principal coordinate analysis of 24 *P. amaryllifolius* accessions based on two EST-SSR markers.

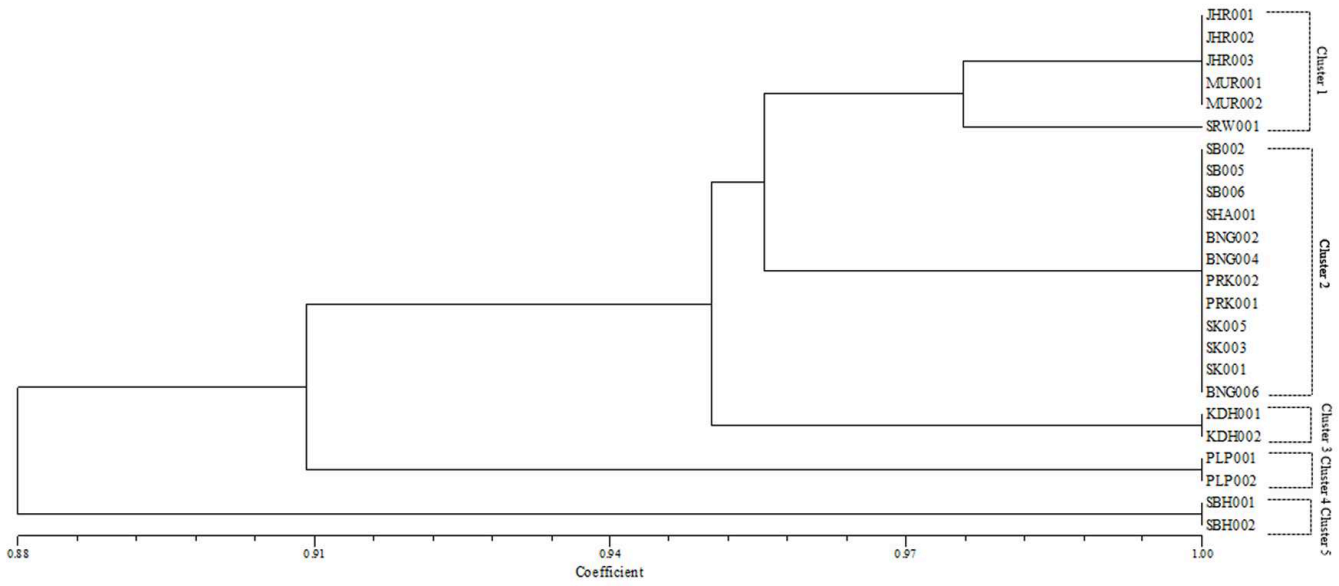


Figure 9

UPGMA dendrogram of 24 P. amaryllifolius accessions based on three iPBS markers.

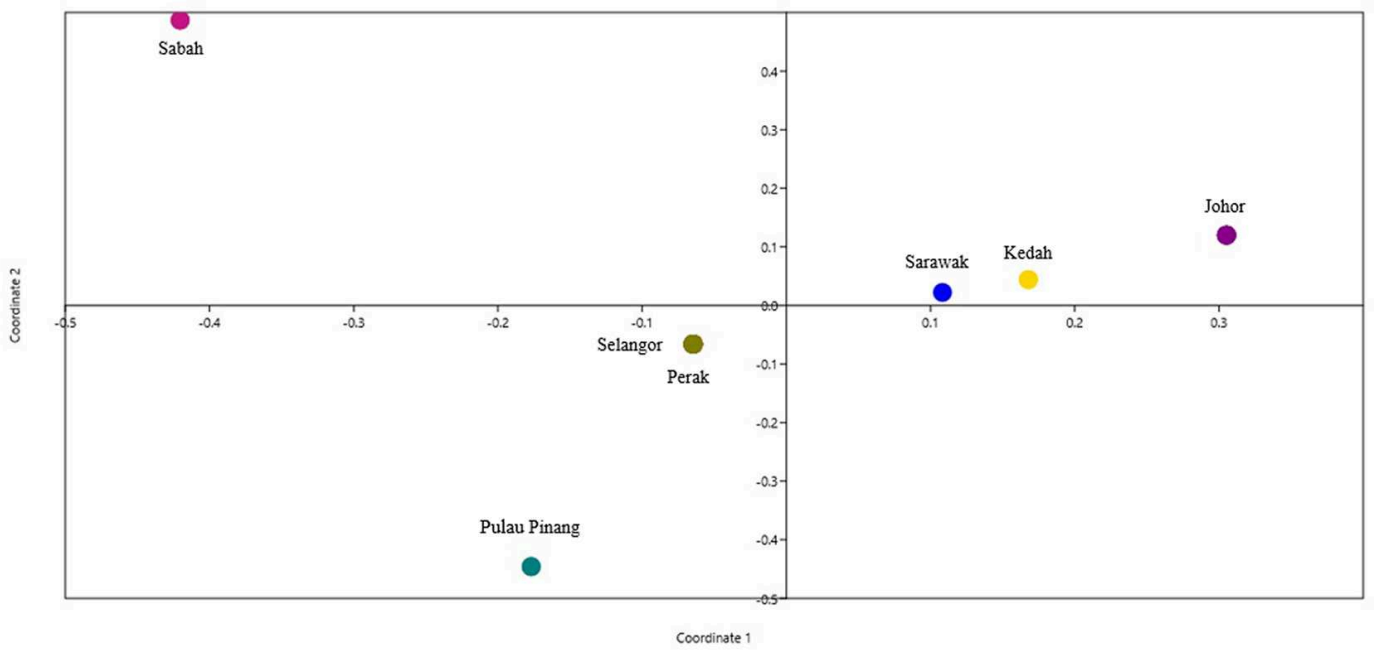


Figure 10

Principal coordinate analysis of 24 *P. amaryllifolius* accessions based on three iPBS markers.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFile1.xlsx
- Supplementaryfiles.docx
- Table1.docx
- Table2.docx
- Table3.docx
- Table4.docx
- Table5.docx
- Table6.docx