TROPICAL LIFE SCIENCES RESEARCH

Genome-Wide SNP Analysis Reveals Distinct Genetic Clusters and Admixture Patterns in Sago Palm

Authors:

Evra Raunie Ibrahim*, Vijay Subbiah Kumar, Muhammad Safiy Syahmi Jamil, Intan Nur Hannani Abdul Halim

***Correspondence:** vijay@ums.edu.my

Submitted; 24 October 2026; Accepted; 5 August 2026; Early view: 4 November 2026
To cite this article: Evra Raunie Ibrahim, Vijay Kumar Subbiah, Muhammad Safiy Syahmi Jamil and Intan Nur Hannani Abdul Halim. (in press). Genome-wide SNP analysis reveals distinct genetic clusters and admixture patterns in sago palm. *Tropical Life Sciences Research.*

Highlights

- Genome-wide resequencing identified 376,875 high-confidence SNPs across 72 sago palm accessions.

- SNP analyses resolved three major varietal lineages, with ADMIXTURE supporting four ancestry components (K = 4).

  The KR thorny low-yield group was genetically distinct and showed the highest genetic diversity, highlighting its value for breeding and conservation.

© Penerbit Universiti Sains Malaysia 2026

EARLY VIEW

Genome-Wide SNP Analysis Reveals Distinct Genetic Clusters And Admixture Patterns In Sago Palm

[1,2]Evra Raunie Ibrahim*, [1]Vijay Subbiah Kumar, [2]Muhammad Safiy, [2]Intan Nur Hannani

[1]Biotechnology Research Institute, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia.
[2]Craun Research Sdn Bhd, Lot 3147 Blok 14, Jalan Sultan Tengah, 93050 Kuching, Sarawak, Malaysia.

*Corresponding author: vijay@ums.edu.my

Running head: SNP-Based Analysis of Sago Palm

Submitted; 24 October 2026; Accepted; 5 August 2026; Early view: 4 November 2026
To cite this article: Evra Raunie Ibrahim, Vijay Kumar Subbiah, Muhammad Safiy Syahmi Jamil and Intan Nur Hannani Abdul Halim. (in press). Genome-wide SNP analysis reveals distinct genetic clusters and admixture patterns in sago palm. *Tropical Life Sciences Research.*

Abstract. The sago palm (*Metroxylon sagu* Rottb.) is an important starch-producing crop in Sarawak, Malaysia, but its genetic improvement and varietal authentication remain limited by the lack of genome-wide molecular markers. This study applied Illumina-based low-depth whole-genome resequencing to identify single nucleotide polymorphisms (SNPs) and assess genetic differentiation among sago palm varieties. Seventy-two accessions representing thorny high-yield, thornless high-yield and thorny low-yield varieties were sequenced using the Illumina NovaSeq platform. Sequencing generated 4.3–9.7 Gb of data per sample, corresponding to 8.4×–19.1× genome coverage, with an average coverage of 11.7×. Variant calling initially identified 4,129,378 SNPs, of which 376,875 high-confidence SNPs were retained after filtering. The selected SNP dataset had 11.79% missing data, 10.38% heterozygous sites and an average minor allele frequency of 0.0706. Population-level diversity estimates showed that expected heterozygosity was highest in KR (H_e = 0.217), followed by TS (H_e = 0.187), MC (H_e = 0.187), MS (H_e = 0.188), AB (H_e = 0.179) and TC (H_e = 0.176). Genome-wide SNP analyses using

phylogenetic clustering, genetic distance analysis, principal component analysis and ADMIXTURE revealed genetic structuring among the sampled accessions. PCA separated the major varietal groups, with PC1 and PC2 explaining 20.5% and 9.0% of the genetic variation, respectively. ADMIXTURE analysis supported K = 4 as the best-fitting model, indicating three main varietal lineages and an additional ancestry component consistent with substructure or admixture. Thorny high-yield and thornless high-yield accessions generally formed distinct clusters, whereas the thorny low-yield group showed greater genetic heterogeneity and partial ancestry sharing with other groups. These findings demonstrate the utility of genome-wide SNP markers for resolving genetic relationships among sago palm varieties and provide a foundation for germplasm management, varietal authentication and future breeding strategies.

Keywords: *Metroxylon sagu*, Genetic Diversity, Population Structure, Single Nucleotide Polymorphism (SNP), Whole Genome Sequencing

INTRODUCTION

Sago palm (*Metroxylon sagu* Rottb.) is a vital starch-producing crop in Southeast Asia, where it contributes to food security and rural economies. Unlike major cereals, sago palm thrives in waterlogged and marginal soils, making it an important resource for sustainable agriculture in peatland ecosystems (Flach, 1997; Ehara *et al.*, 2018). Despite its economic and ecological importance, sago palm remains underutilized and poorly studied at the genomic level compared to other palms such as oil palm (*Elaeis guineensis*) and coconut (*Cocos nucifera*).

Low-depth whole-genome resequencing, also referred to as low-coverage whole-genome sequencing (lcWGS), has emerged as a cost-effective reduced-depth sequencing strategy for large-scale genomic studies, enabling the discovery of single nucleotide polymorphisms (SNPs) and providing insights into population structure, genetic diversity and evolutionary history. Although lower sequencing depth may introduce challenges such as missing data and reduced power to detect rare variants, these limitations can be mitigated through genotype likelihood, imputation and statistical approaches, making lcWGS increasingly applied in conservation genomics, evolutionary biology and crop improvement programs (Lou *et al*., 2021; Baraja-Fonseca *et al*. 2025)

In palm species, genome-wide sequencing and molecular marker approaches have facilitated the identification of markers associated with yield, fruit morphology and disease resistance, while also improving understanding of population structure, genetic diversity and adaptation across diverse environments (Singh *et al*., 2013; Riangwong *et al*., 2020). However, genomic characterization in sago palm remains limited. Although adaptive traits such as waterlogging tolerance and starch variability have been described, their genetic basis is still poorly understood. Furthermore, the genetic status of germplasm collections in Sarawak remains debated, with some studies suggesting distinct varietal groupings whereas others indicate admixture or substantial within-group diversity. This uncertainty highlights the need for molecular data to clarify varietal relationships and support breeding and conservation strategies.

This study addresses these gaps by applying low-coverage whole-genome sequencing (lcWGS) and SNP detection in sago palm. We aimed to generate genome-wide molecular markers, assess genetic variation within and among varieties, and determine the population structure of sago germplasm using multilayered genomic analyses. By integrating genomic data with existing phenotypic knowledge, this study provides a foundation for the genetic improvement, germplasm management and conservation of this underutilized yet economically important palm.

MATERIALS AND METHODS

Plant Materials

Seventy-two young leaf samples of *sago* were collected from the CRAUN Sago Germplasm Plot in Sungai Talau, Mukah, Sarawak. Samples were grouped into six categories (KR, TS, TC, AB, MS, MC) of 12 individuals each, representing three varieties (T: thorny high-yield, M: thornless high-yield, K: thorny low-yield), including normal, clonal, and abnormal phenotypes (AB).

DNA Extraction and Sequencing

Genomic DNA was isolated using a modified CTAB protocol (Doyle and Doyle, 1987) optimized for sago palm, incorporating Chelex® resin to remove contaminants (Walsh *et al.*, 1991). DNA quality was assessed via agarose gel electrophoresis and spectrophotometry, and

samples with OD260/280 between 1.8–2.0 and OD260/230 ≥ 2 were used for sequencing. Libraries for each sample (≥10 µg DNA) were prepared with the TruSeq PCR-free kit (Illumina) using 350–550 bp inserts. Sequencing was conducted on the Illumina NovaSeq 6000 platform (paired-end 150 bp) with service provided by Macrogen Inc., Singapore.

Bioinformatics Workflow

Raw sequencing reads were assessed for quality using FastQC (Andrews, 2010) and trimmed using Trimmomatic (Bolger *et al*., 2014). Clean reads were aligned to the *Metroxylon sagu* reference genome (Lim *et al*., 2022) using BWA-MEM (Li, 2013). Alignment files were sorted, indexed and processed for duplicate removal using SAMtools (Li *et al*., 2011). Variant calling was performed using BCFtools v1.16 (Danecek *et al*., 2021), and initial SNP filtering retained variants with a minimum quality score (QUAL) of 30 and minimum read depth (DP) of 30. Additional filtering criteria based on mapping quality bias (MQBZ), read position bias (RPBZ) and strand bias were also applied to improve variant reliability.

The resulting SNP dataset was further refined by removing monomorphic loci, SNPs with more than 80% missing data and variants with a minor allele frequency (MAF) below 0.01. To reduce linkage among closely adjacent variants, SNP density was thinned by retaining only one SNP within 1 kb intervals. The final filtered SNP dataset was used for downstream population genomic analyses.

Genetic and Population Analysis

Processed SNP data were filtered to remove monomorphic sites, those with >80% missing data, and clustered sites (<1 kb apart). Phylogenetic relationships using the Neighbor-Joining method and principal component analysis (PCA) were performed (Patterson, 2006) using TASSEL v5.2.81 (Bradbury *et al*., 2007), while population structure analysis was conducted using ADMIXTURE (Alexander *et al*., 2009). Data visualization and statistical analyses were performed in R and RStudio (R Core Team, 2023; RStudio Team, 2020). The flow bioinformatic workflow is summarized in Figure 1.

Population genetic diversity parameters were estimated from the selected SNP dataset comprising 376,875 high-confidence SNPs. Accessions were grouped according to the six

sample groups: AB, KR, MC, MS, TC and TS. Observed heterozygosity (Ho), expected heterozygosity (He) and the inbreeding coefficient (F_{IS}) were calculated using the hierfstat package in R (Goudet, 2005). Individual-level H_o and H_e values were first estimated for each accession, followed by population-level summarisation based on the mean and standard deviation within each group. F_{IS} was calculated as $1 - H_o/H_e$ using population mean values to assess heterozygote deficiency or excess within each group.

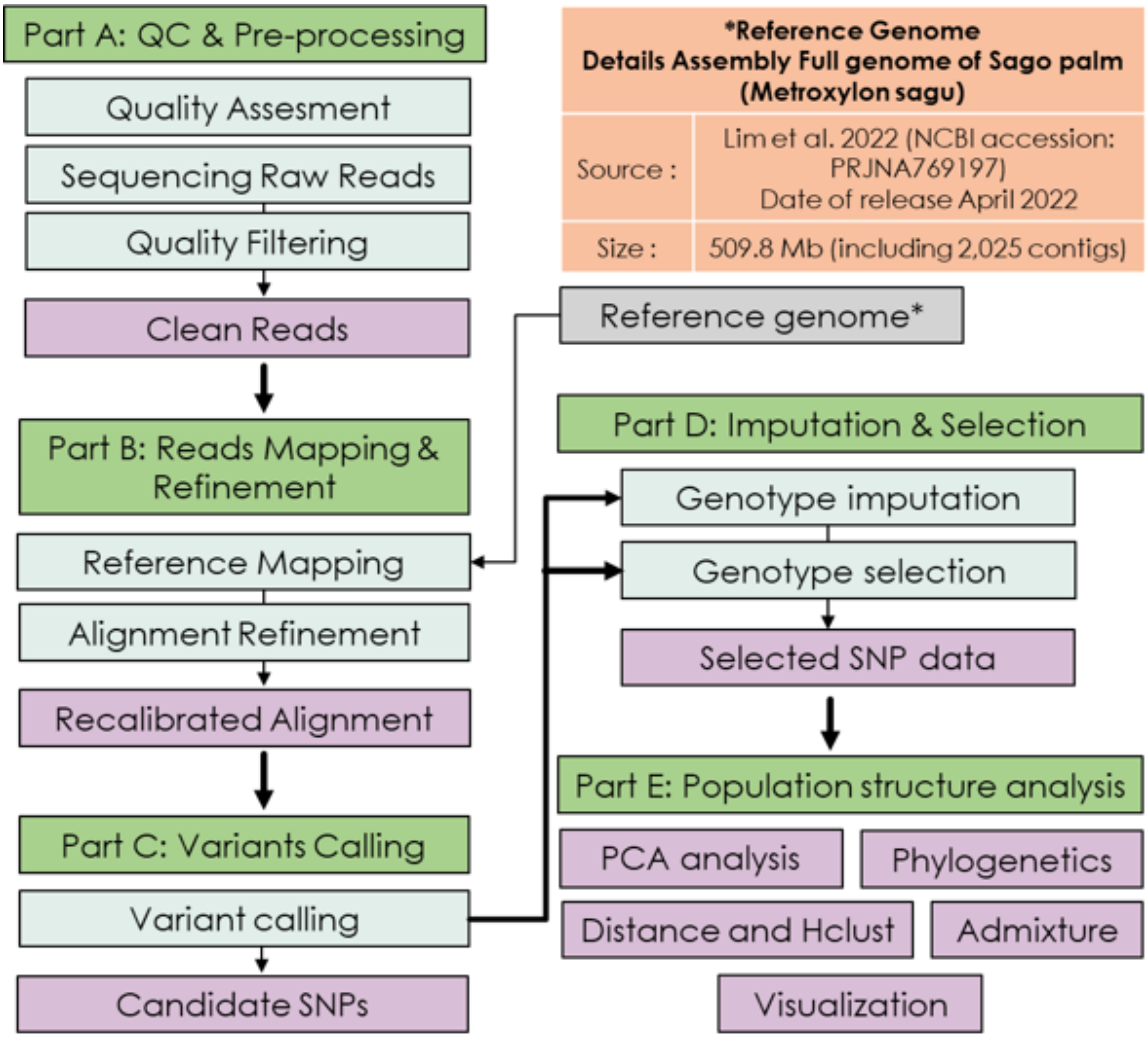


Figure 1. Data analysis pipeline for raw data gained from low coverage whole genome sequencing. At the first stage, the raw data was put through quality assessment and filtering to produce clean reads.

RESULTS AND DISCUSSIONS

Plant Materials and Genomic DNA Preparation

The sago palm accessions were classified into three main varieties (KR, T, and M) according to the CRAUN Sago Germplasm Database (CRAUN Research Sdn. Bhd., unpublished data). The KR variety represents thorny, short-maturing palms with low starch yield. The T variety comprises thorny, long-maturing, high-yielding palms, represented by TS (suckers-derived), TC (clonal), and AB (abnormal clones). TS and TC palms generally have higher starch yield, while AB palms are abnormal clones that show solitary, stunted and non-trunking characteristics. The M variety consists of thornless, long-maturing, high-yielding palms, represented by MS (suckers-derived) and MC (clonal).

DNA quality and concentration were assessed using spectrophotometry (NanoDrop, Thermo Fisher Scientific) and fluorometry (Qubit, Thermo Fisher Scientific), with OD260/280 values of 1.8–2.0 and OD260/230 ≥ 2.0, meeting Illumina library preparation requirements. DNA quantity averaged ≥1.6 µg, exceeding the minimum threshold of 800 ng.

Sequencing Output and coverage

The high-throughput sequencing of 72 samples was successfully carried out using the Illumina NovaSeq platform with paired-end 150 bp (PE150) reads. This approach yielded between 4.3 and 9.7 gigabases (Gb) of data per sample, with an average yield of 6.0 Gb. Such data output is consistent with the expected performance of the Illumina platform system, which is widely recognized for its scalability, throughput, and accuracy (Bentley *et al.*, 2008; Mardis, 2017). The sequencing coverage across samples ranged from 8.4x to 19.1x, averaging 11.7x. This depth is generally sufficient for variant discovery and genotyping in population-scale studies, particularly for detecting common and moderately frequent variants (Liu *et al.,* 2012; Sims *et al.*, 2014). Although higher coverage may be required for rare variant detection, the achieved depth provides a cost-effective balance between sequencing resolution and efficiency.

All sequencing data were generated in raw FASTQ format, preserving nucleotide sequences and associated quality scores. This format enables downstream preprocessing steps such as adapter trimming, quality filtering, and read alignment (Cock *et al.,* 2010). Overall, the sequencing output provides a robust dataset for downstream genomic analyses.

Sequence Raw Data QC and Pre-Processing

The raw sequencing generated 39.8 million reads (5.98 billion bases). After quality control steps such as adapter trimming, filtering of low-quality or ambiguous reads, and removal of short sequences, the data were reduced to 38.1 million reads (5.50 billion bases). In total, about 1.7 million reads (4.3%) and 472 million bases (7.9%) were discarded during preprocessing. The reduction in read number after preprocessing is expected and reflects the removal of low-quality or contaminant sequences that could otherwise compromise downstream analyses such as alignment, SNP calling, and genotyping (Del Fabbro *et al.*, 2013; Bolger *et al.*, 2014). Such artifacts often arise from sequencing errors, particularly toward the 3′ ends of reads, or from degraded DNA templates (Schmieder & Edwards, 2011). By trimming and filtering these unreliable regions, only high-confidence reads are retained, producing a cleaner dataset that improves the accuracy of subsequent analyses. In this study, over 92% of total bases were preserved, indicating that the sequencing run was of high quality with minimal technical artifacts or contamination. Overall, preprocessing is a critical step in NGS workflows, ensuring robust and biologically meaningful data for genetic analysis (Andrews, 2010; Lindgreen, 2012).

Reads Mapping and Refinement

The mapping statistics of clean reads across all samples revealed three categories: uniquely mapped, multiply mapped, and unmapped reads (Figure 2). Most reads were uniquely mapped, reflecting successful alignment to the *M. sagu* draft genome (Lim *et al.*, 2022) and overall high sequencing quality. A smaller proportion of reads were multiply mapped, consistent with repetitive or homologous regions within the genome, while unmapped reads varied slightly among samples but generally remained within expected ranges. Samples such as AB1, KR5, and TC21 showed higher proportions of uniquely mapped reads, indicating good sequencing and alignment quality, whereas samples like MS13 and MS19 displayed more unmapped reads, potentially due to sequence divergence, residual contamination, or read quality issues.

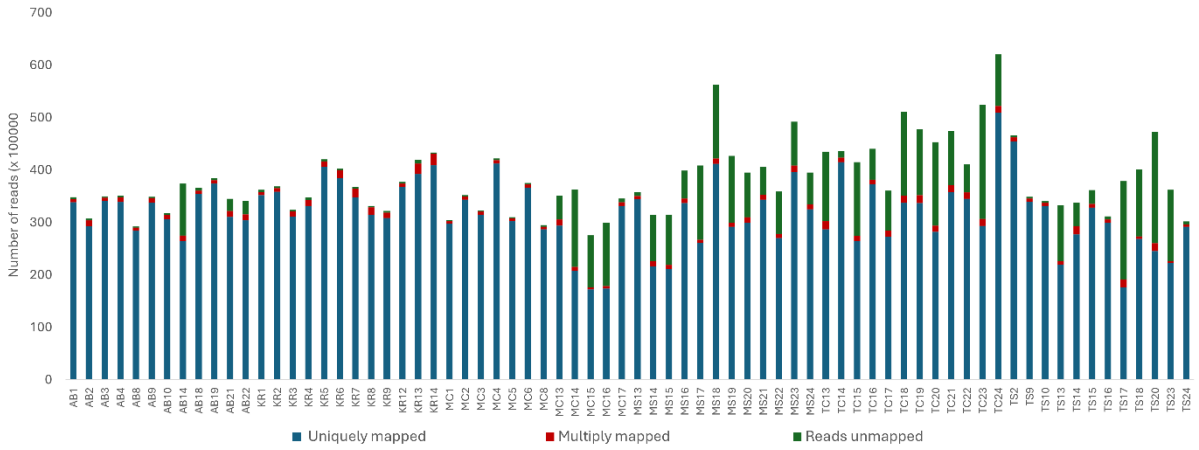


Figure 2. Read mapping statistics for six groups (AB, KR, MC, MS, TC, TS). Bars show the number of reads (×10⁵) that were uniquely mapped (blue), multiply mapped (red), or unmapped (green). Most reads were uniquely mapped, with variation in mapping efficiency observed among groups.

On average, 87% of reads aligned to the reference genome, covering approximately 88% of the genome assembly. This high mapping rate supports the reliability of the dataset for downstream analyses, including variant calling and genotyping. However, the presence of unmapped reads may largely reflect the fragmented state of the current sago palm genome assembly. Although *M. sagu* is diploid (2n = 26; Verhaar, 1976), the available reference remains incomplete, consisting of 2,025 contigs rather than chromosome-scale scaffolds (Lim *et al.*, 2022). Such fragmentation, often caused by unresolved repetitive elements or assembly limitations, is known to reduce alignment efficiency, as certain genomic regions are absent or poorly represented in draft assemblies (Treangen & Salzberg, 2012). Therefore, the observed proportion of unmapped reads may largely reflect reference genome incompleteness than contamination or sample quality issues.

Variants Calling and SNP Filtering

From the recalibrated alignment data, single nucleotide variants (SNVs) were jointly called across all samples, yielding 4,129,378 filtered variants after removal of probable false positives. Prior to downstream analyses, the dataset was further refined to retain 376,875 high-quality variants through stringent filtering based on missing data thresholds, minor allele frequency (MAF), and heterozygosity levels. These criteria reduced the dataset substantially,

ensuring reliability for population structure and relatedness analyses (Anderson *et al.*, 2010; Purcell *et al.*, 2007). The summary statistics of both filtered and selected variants are presented in Table 1.

Table 1: The variant statistics of the resulting genotypes for 72 samples.

Statistic	Filtered variants	Selected variants
Number of individuals (taxa)	72	72
Number of SNP sites	4,129,378	376,875
Total genotype calls	297,315,216	27,135,000
Non-missing genotype calls	263,289,147	23,937,000
Proportion non-missing	0.8856	0.8822
Missing genotype calls	34,026,069	3,197,900
Proportion missing	0.1144	0.1179
Heterozygous genotype calls (summed across individuals)	36,444,112	2,817,100
Proportion heterozygous	0.1226	0.1038
Mean minor allele frequency (MAF)	0.0808	0.0706

For the filtered dataset, 11.44% of genotype calls were missing (34,026,069 out of 297,315,216 total genotype calls), while the selected dataset showed a comparable proportion of missing data at 11.79% (3,197,900 out of 27,135,000 genotype calls). This consistency in missingness across datasets indicates that the filtering process maintained overall data quality while reducing the number of loci, rather than introducing systematic loss of information (Li *et al*., 2011).

Heterozygosity decreased from 12.26% (36,444,112 heterozygous genotype calls) in the filtered dataset to 10.38% (2,817,100 heterozygous genotype calls) in the selected dataset. This reduction suggests that stringent filtering preferentially retained more homozygous loci and reduced genotype uncertainty, which is consistent with standard practices in population genomic analyses where conservative filtering improves robustness of downstream inferences (Purcell et al., 2007). Similarly, the average minor allele frequency (MAF) decreased from 0.0808 in the filtered dataset to 0.0706 in the selected dataset, reflecting the removal of rare

alleles, a common approach to improve the reliability of population structure and diversity analyses (Maruki & Lynch, 2017).

Overall, the filtered and selected datasets maintained high completeness, with approximately 88% of genotype calls retained as non-missing data, providing a high-quality SNP dataset for downstream genetic diversity and population structure analyses (Figure 3).

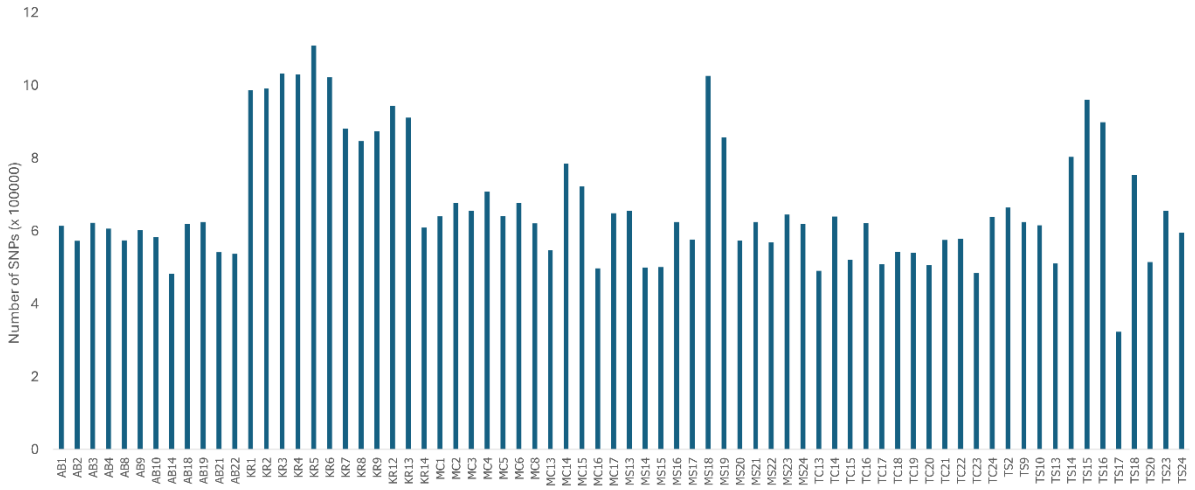


Figure 3. Number of SNPs detected for six groups (AB, KR, MC, MS, TC, TS). SNP counts ranged from ~4 × 10^5 to >11 × 10^5, with KR and MC generally showing higher values and AB and TS lower values.

Samples from the KR group, notably KR1, KR4, KR7, and KR12, exhibited the higher number of called SNPs, which may reflect higher levels of genetic variation, better sequencing coverage, or more complete genotype recovery after filtering.. In contrast, several samples such as AB10, AB19, TC22, TS13, and TS2 showed lower SNP counts, which may be associated with reduced genomic diversity, lower sequencing depth, or stricter filtering of low-quality genotype calls. The remaining samples across the AB, MC, MS, TC, and TS groups displayed moderate and relatively consistent SNP numbers.

The observed variation is consistent with previous findings in population-scale sequencing studies, where both biological such as genetic diversity and population structure) and technical factors (including sequence depth, reference genome quality, and filtering parameters) can strongly influence variant detection (DePristo *et al.*, 2011; Nielsen *et al.*, 2011). Such uneven

distribution of variant calls can influence downstream analyses such as clustering, PCA, and phylogenetic inference, if not properly addressed (Davey *et al.*, 2011).

In this study, the application of stringent filtering ensured that only high-confidence SNPs were retained, minimizing noise while preserving informative variation. Although this process may slightly increase missing data and reduce heterozygosity, such trade-offs are common in population genomics and are necessary for robust genetic diversity and structure analyses (Hickey *et al.*, 2012; Schmidt *et al.*, 2021). The relatively high SNP counts observed in several KR samples also indicate the presence of substantial genetic diversity, highlighting its potential value for future breeding and conservation efforts.

Genetic Relationship and Population Structure Analysis for 72 Sago Samples

Exploring more on the data quality and genetic variation among groups, we summarized the number of missing data points and heterozygosity across the six sago palm groups (AB, KR, MC, MS, TC, TS) (Table 2). While all groups had the same number of sites retained after filtering (376,875 SNPs), clear differences were observed in missing data and heterozygosity levels. KR showed the highest heterozygosity (19.9%), followed by TS (12.6%), whereas AB and TC exhibited lower values (7.7–8.0%). Proportion of missing data also varied, ranging from 9.8% in AB to 14.4% in TS. These differences reflect variation in genetic diversity and data completeness, which may influence subsequent clustering and structure analyses.

Table 2. Summary of SNP dataset statistics across six sago groups. Each value represents the group mean for the SNP sites, showing levels of missing data and heterozygosity.

Group	Number of Sites	Gametes Missing	Proportion Missing	Number Heterozygous	Proportion Heterozygous
AB	376875	73834.83	0.098	27102.1667	0.0796
KR	376875	92431.83	0.1226	65697.8333	0.1985
MC	376875	81899.83	0.1087	37466.5	0.1138
MS	376875	87341.67	0.1159	37753	0.1126
TC	376875	88967.67	0.118	25747.5	0.0772
TS	376875	108506.2	0.144	40995.25	0.1262

The observed differences in heterozygosity and missing data among groups suggest varying levels of genetic diversity, which are expected to influence patterns of relatedness and

clustering. To visualize these relationships, we next constructed a dendrogram and performed principal component analysis (PCA) based on the filtered SNP dataset.

Population genetic diversity estimates

Population-level genetic diversity estimates were calculated using the selected SNP dataset of 376,875 high-confidence SNPs. Observed heterozygosity (H_o) ranged from 0.185 in TC to 0.216 in KR, while expected heterozygosity (H_e) ranged from 0.176 in TC to 0.217 in KR (Table 3). KR showed the highest genetic diversity among the six groups, with H_o = 0.216 ± 0.016 and H_e = 0.217 ± 0.016. In contrast, AB and TC showed slightly lower diversity, with H_e values of 0.179 and 0.176, respectively. MC, MS and TS showed intermediate diversity values, with H_e ranging from 0.187 to 0.188.

The inbreeding coefficient (F_{IS}) varied among groups. KR showed a value close to zero (F_{IS} = 0.008), suggesting near-equilibrium heterozygosity levels. The remaining groups showed negative F_{IS} values, ranging from −0.058 in MS to −0.040 in TS, indicating an excess of heterozygotes relative to expectation. This pattern may reflect outcrossing, admixture or the mixed genetic background of germplasm accessions. Overall, the H_o, H_e and F_{IS} estimates support the clustering analyses by showing that KR is the most genetically diverse group, whereas AB and TC show comparatively lower diversity.

Table 3. Population genetic diversity parameters of six *Metroxylon sagu* groups based on 376,875 high-confidence SNPs.

Population	H_o (mean ± SD)	H_e (mean ± SD)	F_{IS}
AB	0.188 ± 0.003	0.179 ± 0.003	-0.051
KR	0.216 ± 0.016	0.217 ± 0.016	0.008
MC	0.196 ± 0.014	0.187 ± 0.013	-0.047
MS	0.199 ± 0.028	0.188 ± 0.022	-0.058
TC	0.185 ± 0.007	0.176 ± 0.006	-0.049
TS	0.194 ± 0.021	0.187 ± 0.020	-0.040

Phylogenetic tree analysis

A phylogenetic tree constructed from SNP data revealed three major clusters among the 72 *M. sagu* accessions (Figure 4). Cluster 1 grouped the TS, TC, and AB accessions, showing short branch lengths and indicating high genetic similarity. This suggests a shared genetic background, consistent with their origin from the same variety, although the clustering did not clearly distinguish between sucker-derived, clonal, and abnormal clonal lines.

Cluster 2 comprised the MS and MC accessions, which displayed broader genetic divergence compared to Cluster 1. This group represents the thornless variants, and the observed variation may reflect differences in propagation practices or local adaptation (Teh *et al.*, 2011; Tumonggor *et al.*, 2013).

Cluster 3 contained the KR accessions, which were genetically distinct from the other groups. This separation highlights significant divergence, possibly shaped by geographic isolation or different selective pressures, suggesting that the KR population may require independent conservation or breeding strategies (Clement *et al.*, 2017).

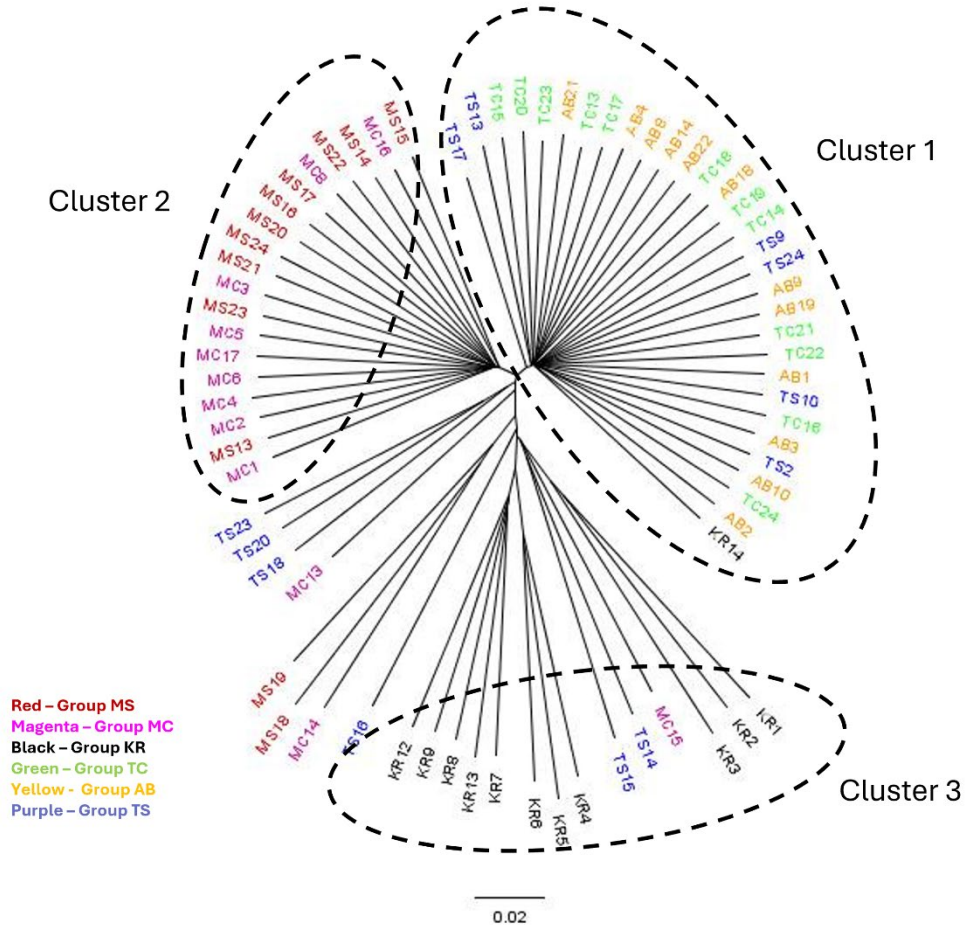


Figure 4. Based on clustering analysis, the Phylogenetic dendrogram illustrates the genetic relationships among samples T, M and KR group. Samples cluster according to thorny (TS, TC, AB) and thornless (MS, MC) groups, with KR forming a distinct subgroup.

A few samples (TS18, TS14, TS15, TS16, MS18, and MC14) were positioned outside their expected genetic clusters. Such outlier placement may arise from several biological processes. First, gene flow and introgression between cultivated populations could result in genetic signals that do not align with phenotypic classifications (Slatkin, 1987). Second, admixture or hybridization can generate individuals with intermediate ancestry, producing intermediate or unexpected clustering patterns (Pritchard *et al.*, 2000). In tissue culture–derived palms such as MC14, somaclonal variation is also a likely contributor, as *in vitro* propagation can introduce novel mutations and epigenetic changes that differentiate regenerated clones from their source (Bairu *et al.*, 2011). Finally, some outliers may reflect cryptic genetic diversity or potential misclassification of samples, given that phenotypic traits such as thorniness may not fully capture the underlying genetic relationships (Allendorf *et al.*,

2013). Together, these factors highlight the complex evolutionary and propagation history of sago palm populations.

Overall, the SNP-based phylogeny reflects population structure shaped by both clonal and seed-based propagation, supporting the utility of SNP markers for resolving genetic relationships in sago palm. The distinct clustering of KR accessions underscores their unique genetic identity, which should be considered in future breeding and conservation programs.

Distance analysis based on selected SNPs

Hierarchical clustering of 72 *M. sagu* accessions based on SNP-derived genetic distances revealed clear subgrouping consistent with the phylogenetic tree (Figure 5). Three primary clusters were observed. The first major cluster contained the MC and MS accessions, which grouped closely together, suggesting a recent shared ancestry or exposure to similar selective pressures. The second cluster comprised the KR accessions, which formed a tight and genetically distinct group, reflecting divergence likely shaped by geographic isolation or limited gene flow. The third and more heterogeneous cluster included AB, TC, and TS accessions. These samples were dispersed across sub-branches, suggesting admixture, shared ancestry, or exchange of planting materials across populations.

The observed clustering patterns align with expectations for geographically or ecologically structured populations, where reduced gene flow and local adaptation promote divergence (Pritchard *et al.*, 2000; Excoffier & Heckel, 2006). The close grouping of MC and MS may reflect their geographical proximity, while the distinctiveness of KR highlights its potential as a genetically unique population. Notably, some TS and AB accessions were positioned within clusters dominated by other groups, suggesting historical admixture or recent hybridization. Similar cases of outlier placement have been reported in other clonally propagated crops such as cassava (*Manihot esculenta*) (Bhattacharjee *et al.*, 2012) and also banana (*Musa* spp.), where AFLP-based studies and tissue culture-derived variants revealed divergent genotypes within otherwise uniform populations (Bairu *et al.*, 2011). Such outlier patterns are often attributed to somaclonal variation, spontaneous mutations, or historical introgression events, all of which can contribute to unexpected genetic differentiation in clonally propagated crops (Bairu and Kane., 2011).

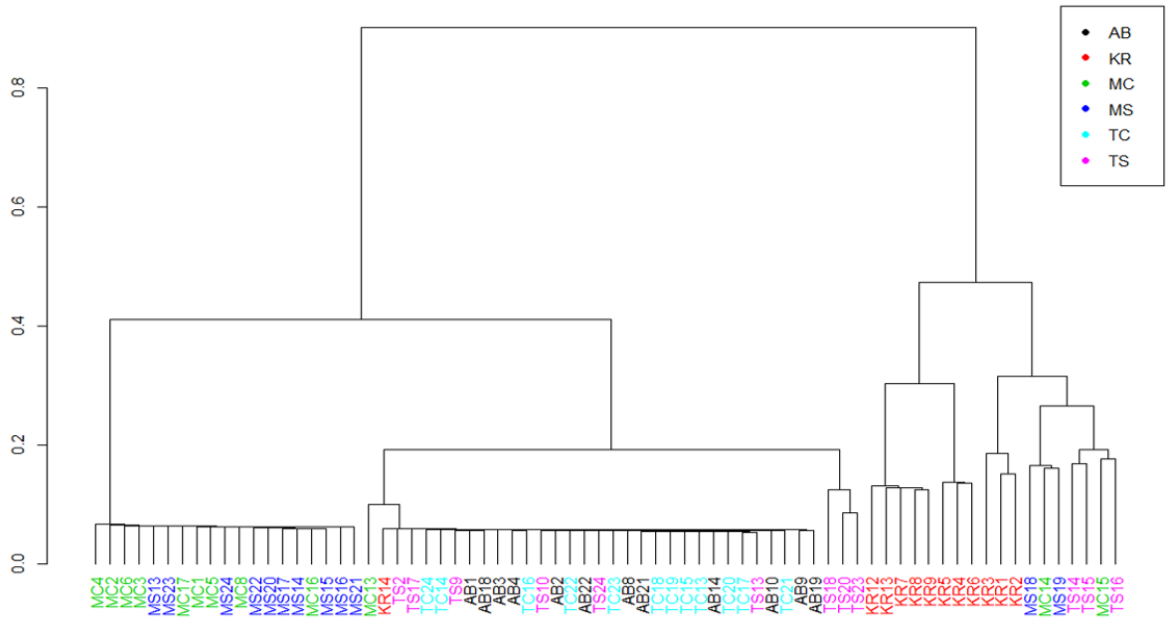


Figure 5. Dendrogram showing genetic relationships based on SNP data. Samples cluster according to thorny (TS, TC, AB) and thornless (MS, MC) groups, with KR forming a distinct subgroup.

Overall, the SNP-based dendrogram provided strong resolution of genetic relationships and diversity patterns in *M. sagu*. The clear separation of KR, MS/MC, and the mixed AB/TC/TS group supports the presence of distinct evolutionary lineages. These results emphasize the effectiveness of genome-wide SNP markers in resolving genetic differentiation and highlight the need to conserve the unique KR gene pool while recognizing ongoing gene flow among cultivated populations.

Heat map analysis of genetic relationships

To complement the dendrogram results, a heat map of pairwise genetic distances was generated from SNP data for 72 *Metroxylon sagu* accessions. The heat map (Figure 6) revealed distinct clustering patterns consistent with dendrogram and phylogenetic analyses, while also providing a continuous visualization of subtle relatedness among individuals.

A prominent cluster in the upper-left quadrant comprised MS18, TS15, MC15, KR3, and KR1, which showed strong similarity (dark red hues) despite belonging to different varietal groups. This unexpected proximity may reflect shared ancestry or historical admixture through planting

material exchange. In contrast, TC and AB samples formed a tight group but with lighter shades, indicating greater internal diversity. Several TS individuals were more scattered, displaying intermediate relationships that suggest admixture or transitional genetic links between populations.

Overall, the heat map supports the separation of KR/MS/MC from TC/AB groups while highlighting the bridging role of TS accessions. These patterns align with previous reports in clonally propagated crops where localized reproduction, mixed cultivation, and human-mediated exchanges shape population structure (Bhattacharjee *et al.*, 2012; Lebot, 2009; Mutegi *et al.*, 2011; Zhao *et al.*, 2015). Together with dendrogram, PCA, and Admixture analyses, the heat map provides robust evidence of genetic structure in *M. sagu* and valuable insights for germplasm management and breeding.

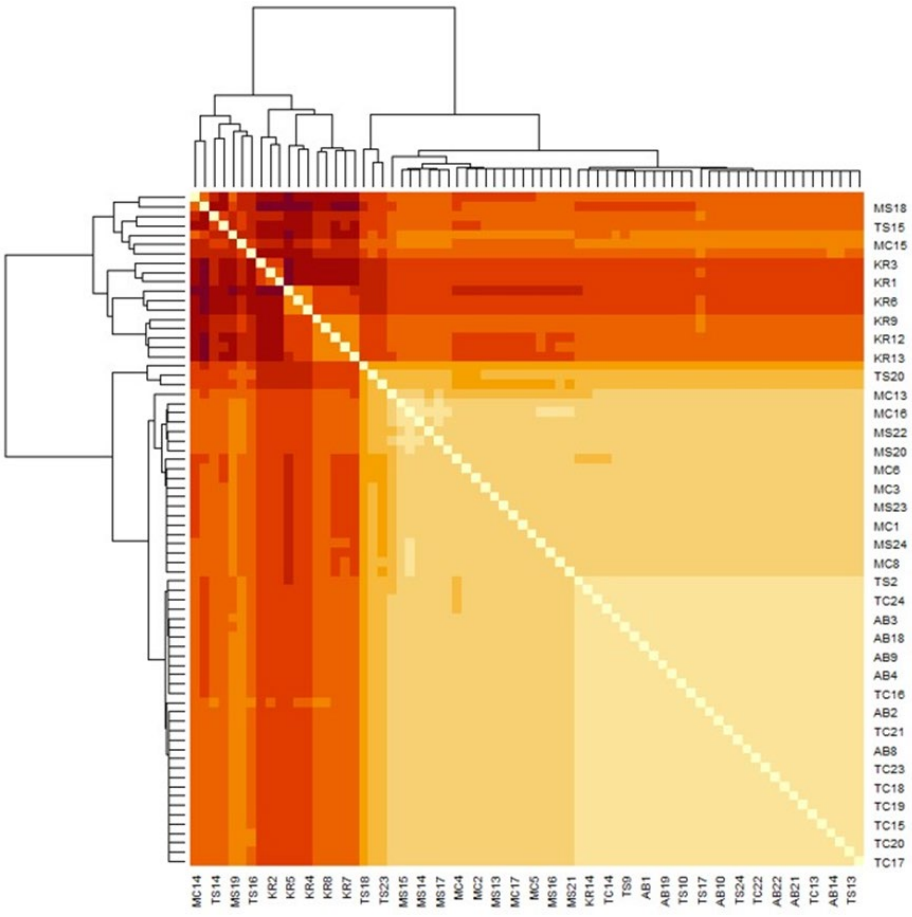


Figure 6. Heatmap of pairwise comparison matrix derived from differential SNP genotypes in Group TS, TC, AB, MS, MC and KR. Heatmap of pairwise genetic distances among 72 sago

palm samples from group TS, TC, AB, MS and MC. Warmer colors indicate higher dissimilarity, while cooler colors indicate greater similarity, reflecting group-specific clustering patterns.

PCA analysis

Principal Component Analysis (PCA) was used to visualize genetic structure among 72 *M. sagu* individuals representing six populations (MS, MC, TC, TS, AB, KR). The scree and cumulative variance plots (Figure 7) indicated that the first two principal components captured the largest proportion of variance, with eigenvalues at PC1 and PC2 explaining 20.5% and 9.0% of the total genetic variation, respectively. Variance accumulation plateaued after the fourth component, suggesting that a reduced set of PCs was sufficient to summarize the data structure, consistent with common patterns observed in SNP-based studies (Jolliffe & Cadima, 2016; Ringnér, 2008).

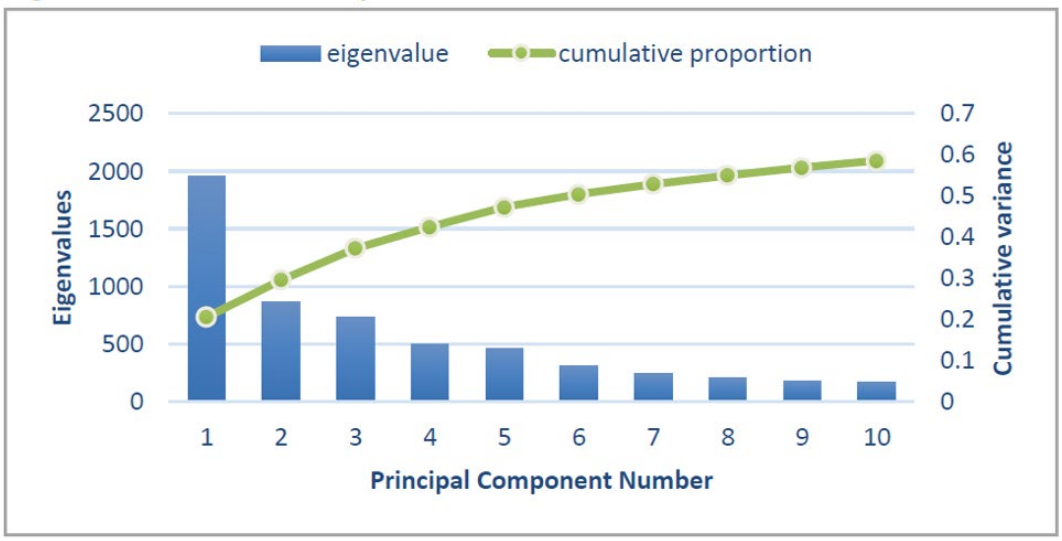


Figure 7. PCA scree plot showing the proportion of genetic variation explained by each principal component (PC) and the cumulative variance across all PCs based on SNP data from 72 sago palm samples. PC1 and PC2 values, together accounting for the total variation.

The scatter plot (Figure 8) revealed three major genetic clusters. MS and TC individuals grouped tightly in the upper-right quadrant, reflecting strong genetic similarity and possibly

recent shared ancestry or geographic connectivity.

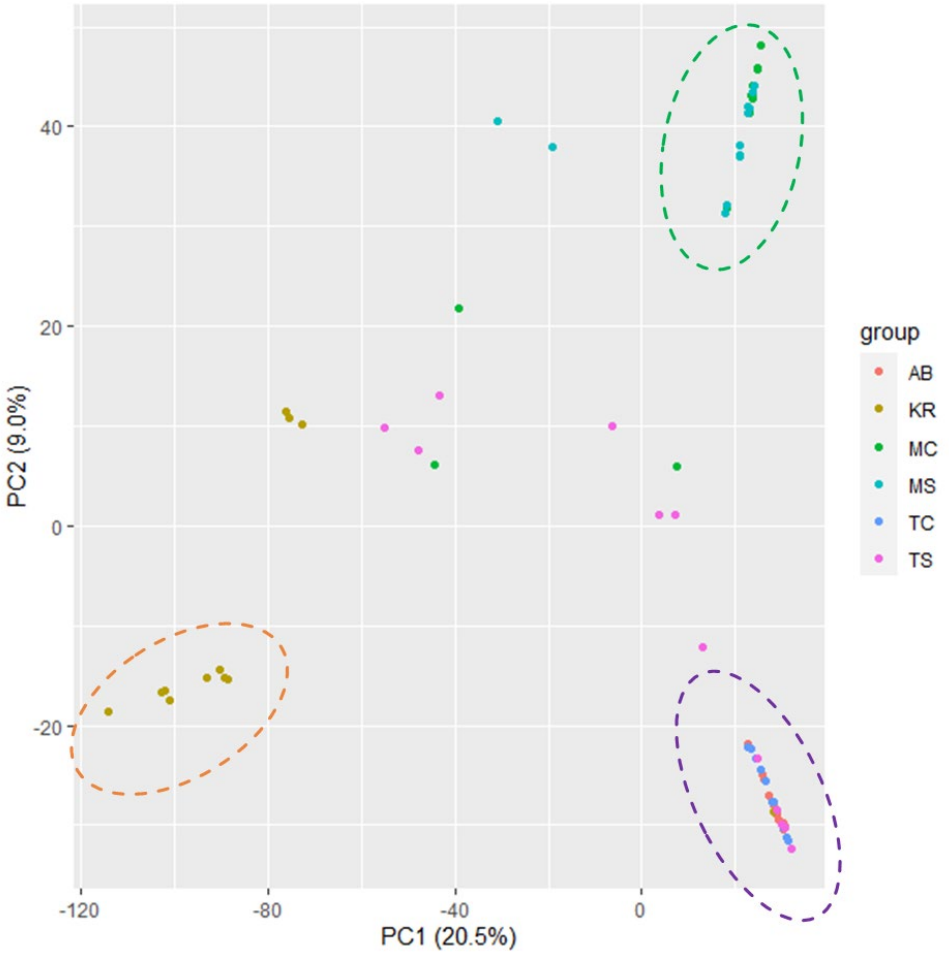


Figure 8. Principal component analysis (PCA) of 72 sago samples based on SNP variation. The first two principal components separate thorny (TS, TC, AB) and thornless (MS, MC) groups, while KR forms a distinct cluster.

In contrast, KR samples were distinctly separated along PC1 in the lower-left quadrant, indicating pronounced genetic divergence, likely due to isolation or population-specific evolutionary pressures. The remaining groups (TS, AB, and MC) displayed scattered distributions between the main clusters, suggesting higher within-population variability and evidence of admixture. Their intermediate positioning is consistent with admixture results at K = 4, which indicated shared ancestry components and possible historical gene flow among these populations.

Overall, PCA results corroborated the clustering patterns observed in the dendrogram, heat map, and admixture analyses, highlighting both distinct population differentiation (KR, MS, TC) and transitional genetic structures (AB, MC, TS). These findings emphasize the genetic complexity of *M. sagu* populations and provide insights into gene flow, population structure, and germplasm management strategies.

Admixture analysis

Admixture analysis was performed on the SNP dataset to determine the number of genetic clusters (K) among the 72 sago samples. Cross-validation (CV) errors (Figure 9) indicated that K = 4 provided the best fit, with the lowest error (0.21), while both lower (K = 2) and higher (K ≥ 5) values showed poorer model performance. The ancestry proportions for K = 2–10 are shown in Figure 10.

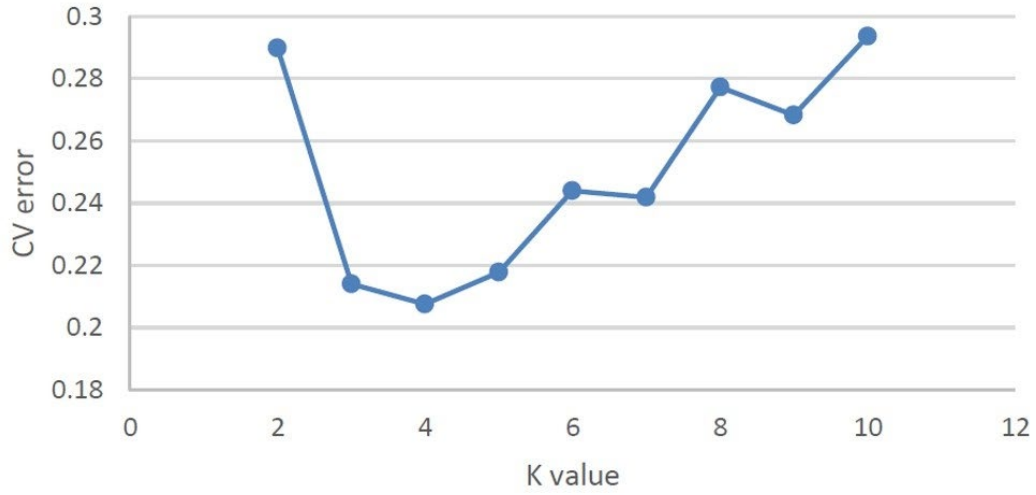


Figure 9. Cross-validation (CV) error values for Admixture analysis across a range of K values (number of genetic clusters). The lowest CV error was observed at K = 4, indicating this as the most likely number of genetic clusters in the sago dataset.

At K = 4, four ancestral clusters were identified, largely corresponding with the three known phenotypic varieties. Variety 1 (TS, TC, and AB) grouped mainly within the same genetic cluster, with AB showing strong similarity to TC despite its abnormal phenotype. Variety 2 (MS and MC) formed a distinct cluster, clearly separated from the thorny varieties. Variety 3 (KR) was genetically isolated, represented by a unique ancestry cluster. A fourth minor cluster was shared in low proportions across groups, suggesting hidden substructure or past gene flow.

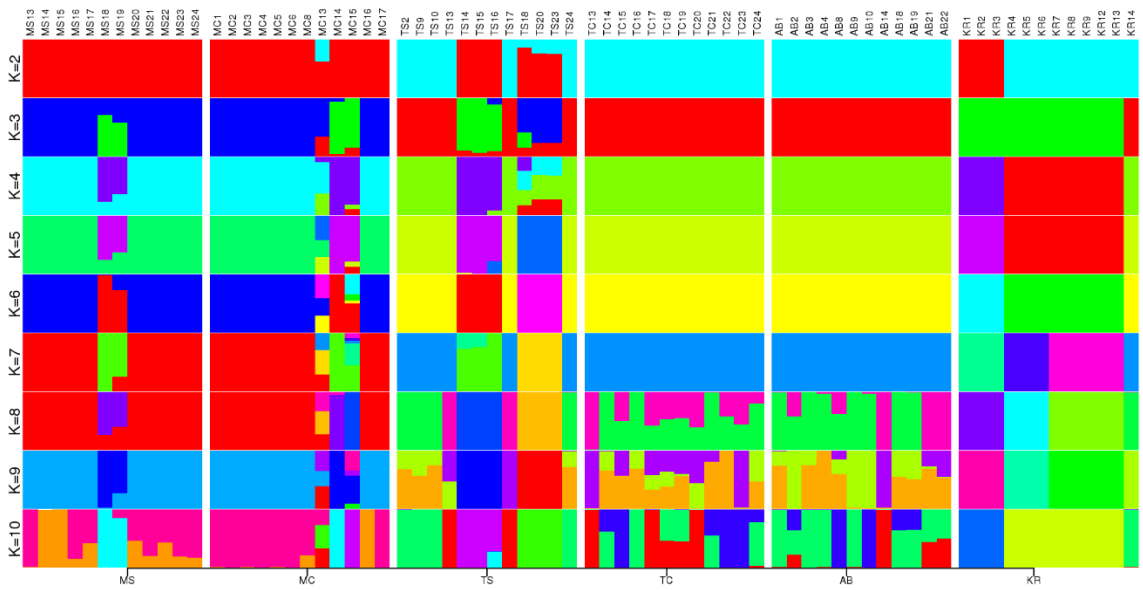


Figure 10. Admixture plot at K = 2 to K = 10 for selected SNPs. The plot showing ancestry proportions of 72 sago accessions at the optimal number of clusters (K = 4). Each vertical bar represents an individual, with colors indicating the proportion of ancestry assigned to each cluster.

The population genetic statistics provided quantitative support for the clustering patterns observed in the SNP-based analyses. Among the six groups, KR showed the highest diversity, with H_o = 0.216 ± 0.016 and H_e = 0.217 ± 0.016, while TC and AB showed lower expected heterozygosity, with H_e = 0.176 and 0.179, respectively. This supports the interpretation that KR represents a more diverse genetic group, whereas TC and AB are comparatively more uniform, possibly reflecting clonal propagation, selection or a narrower genetic background. The F_{IS} values were close to zero or negative across most groups, suggesting no strong heterozygote deficiency and indicating possible outcrossing, mixed ancestry or admixture within the germplasm collection. These diversity estimates strengthen the PCA and ADMIXTURE results by showing that the sampled groups differ not only in clustering pattern but also in within-group genetic diversity.

These results demonstrate that genome-wide SNP markers effectively capture the genetic structure of *M. sagu*, revealing three major varietal lineages and one additional background cluster.

CONCLUSION

Low-coverage whole-genome sequencing combined with SNP detection proved effective in characterizing the genetic diversity of sago palm. The analyses revealed clear varietal differentiation alongside evidence of shared ancestry and admixture among certain groups. PCA, phylogenetic clustering, and admixture consistently supported the population structure, highlighting both distinct lineages and overlaps likely related to hybridization or clonal propagation. This genomic framework provides a reliable basis for varietal identification and offers valuable insights to guide breeding, conservation, and sustainable utilization of sago palm.

FUNDING

This research was funded by the Universiti Malaysia Sabah (UMS) Great Grant Scheme (GUG0426-2/2019) and the Sarawak Government through Craun Research Sdn. Bhd. under the 12th Malaysia Plan funding programme.

ACKNOWLEDGMENTS

The authors sincerely thank the plantation and Biotechnology Laboratory teams of Craun Research Sdn. Bhd. for their technical assistance and for providing plant materials essential to this study. The authors also acknowledge the Biotechnology Research Institute, Universiti Malaysia Sabah, for providing laboratory facilities and research support.

AUTHORS' CONTRIBUTIONS

Evra Raunie Ibrahim: Conceptualization, Investigation, Formal analysis, Writing – original draft, Writing – review and editing.
Muhammad Safiy Syahmi Jamil and Intan Nur Hannani: Investigation, Technical support.
Vijay Subbiah Kumar: Funding, Supervision, conceptualization, Writing – review and editing.

All authors have read and approved the final version of the manuscript.

REFERENCES

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc

Bairu, M. W., Aremu, A. O., & van Staden, J. (2011). Somaclonal variation in plants: Causes and detection methods. *Plant Growth Regulation, 63*(2), 147–173. https://doi.org/10.1007/s10725-010-9554-x

Bairu, M. W., & Kane, M. E. (2011). *Physiological and developmental problems encountered by in vitro cultured plants*. Plant Growth Regulation, 63, 101–103. https://doi.org/10.1007/s10725-011-9565-2

Baraja-Fonseca, V., Arrones, A., Vilanova, S. et al. Benchmarking of low coverage sequencing workflows for precision genotyping in eggplant. BMC Plant Biol 25, 1125 (2025). https://doi.org/10.1186/s12870-025-07242-x

Bhattacharjee, R., Ferguson, M., Gedil, M., Dumet, D. & Ingelbrecht, I. (2009). Field collection, preservation and large-scale DNA extraction procedures for cassava (*Manihot esculenta* Crantz.). African Journal of Biotechnology, 8(15), p. 3424-3430. https://hdl.handle.net/10568/90807

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics, 23*(19), 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Danecek P., Bonfield J. K., Liddle J., Marshall J., Ohan V., Pollard M. O., Whitwham A., Keane T.,McCarthy S. A., Davies R. M., Li H. Twelve years of SAMtools and BCFtools, *GigaScience*, 10(2). https://doi.org/10.1093/gigascience/giab008

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., ... & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics, 43(5), 491–498. https://doi.org/10.1038/ng.806

Doyle, J.J., & Doyle, J.L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*, 19, 11–15.

Ehara, H. (2018). Genetic variation and agronomic features of Metroxylon palms in Asia and Pacific. In H. Ehara, Y. Toyoda, & D. V. Johnson (Eds.), The Sago Palm (pp. 45–59). Springer. https://doi.org/10.1007/978-981-10-5269-9_4

Flach, M. (1997) Sago Palm *Metroxylon sagu* Rottb. Promoting the Conservation and Used of Unther Utilized and Neglected Crops. 13. Institute of Plant Genetics and Crop Plant Research, Gatersleben/Internasional Plant Genetic Resources Institute (IPGRI), Rome.

Goudet, J. (2005). hierfstat, a package for R to compute and test hierarchical F-statistics. Molecular Ecology Notes, 5(1), 184–186. https://doi.org/10.1111/j.1471-8286.2004.00828.x

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Lebot, V. (2009). Tropical root and tuber crops: Cassava, sweet potato, yams and aroids. CABI Publishing. https://doi.org/10.1079/9781845934248.0000

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997 https://doi.org/10.48550/arXiv.1303.3997

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2011). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lim, L. W. K., Chung, H. H., Hussain, H., & Gan, H. M. (2022). First high-quality genome assembly data of sago palm (*Metroxylon sagu* Rottboll). *Data in Brief*, 40, 107800. https://doi.org/10.1016/j.dib.2022.107800

Lou, R. N., Jacobs, A., Wilder, A. P., & Therkildsen, N. O. (2021). A beginner's guide to low-coverage whole genome sequencing for population genomics. Molecular ecology, 30(23), 5966–5993. https://doi.org/10.1111/mec.16077.

Mutegi, E., Sagnard, F., Semagn, K., Deu, M., Muraya, M., Kanyenji, B., … de Villiers, S. (2011). Genetic structure and relationships within and between cultivated and wild sorghum (Sorghum bicolor (L.) Moench) in Kenya as revealed by microsatellite markers. Theoretical and Applied Genetics, 122(5), 989–1004. https://doi.org/10.1007/s00122-010-1504-5

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics, 12(6), 443–451. https://doi.org/10.1038/nrg2986

Nisar, M., & Hussain, H. (2022). Assessment of the genetic variations of sago palm *Metroxylon sagu* in three regions of Sarawak, Malaysia using amplified fragment length polymorphism (AFLP) marker. *Chemical and Biological Technologies in Agriculture, 9*, 46. https://doi.org/10.1186/s40538-022-00315-1

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigen analysis. *PLoS Genetics, 2*(12), e190. https://doi.org/10.1371/journal.pgen.0020190

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics, 155(2), 945–959. https://doi.org/10.1093/genetics/155.2.945

R Core Team. (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Riangwong, K., Wanchana, S., Aesomnuk, W. *et al.* (2020). Mining and validation of novel genotyping-by-sequencing (GBS)-based simple sequence repeats (SSRs) and their application for the estimation of the genetic diversity and population structure of coconuts (*Cocos nucifera* L.) in Thailand. *Horticulture Research* 7(156). https://doi.org/10.1038/s41438-020-00374-1

Ringnér, M. (2008) What is principal component analysis?. *Nature Biotechnology* 26, 303–304. https://doi.org/10.1038/nbt0308-303

RStudio Team. (2020). RStudio: Integrated development for R. RStudio, PBC. http://www.rstudio.com/

Schmidt, T. L., Jasper, M.-E., Weeks, A. R., & Hoffmann, A. A. (2021). Unbiased population heterozygosity estimates from genome-wide sequence data. *Methods in Ecology and Evolution*, 12(10), 1888–1898. https://doi.org/10.1111/2041-210X.13659

Singh, R., Ong-Abdullah, M., Low, ET. *et al.* (2013). Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. *Nature* **500**, 335–339 (2013). https://doi.org/10.1038/nature12309

Slatkin, M. (1987). Gene flow and the geographic structure of natural populations. *Science, 236*(4803), 787–792. https://doi.org/10.1126/science.3576198

Walsh, P.S., Metzger, D.A., & Higuchi, R. (1991). Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. *BioTechniques*, **10**(4), 506–513.

Yong, R. C. M., Sobeng, Y., Zaini, F., & Busri, N. (2018). Suitability of peat swamp areas for commercial production of sago palms: The Sarawak experience. In H. Ehara, Y. Toyoda, & D. V. Johnson (Eds.), *Sago Palm*. Springer. https://doi.org/10.1007/978-981-10-5269-9_7

Zhao, Y., Williams, R., Prakash, C. S., & He, G. (2015). Genetic diversity and relationships of cassava (Manihot esculenta Crantz) germplasm revealed using SSR markers. Genetic Resources and Crop Evolution, 62(7), 1221–1230. https://doi.org/10.1007/s10722-015-0217-7