



Faculty of Resource Science and Technology

Genomic Landscape of Sago Palm (*Metroxylon sagu* Rottb.)

Lim Whye Kit Leonard

**Doctor of Philosophy
2022**

Genomic Landscape of Sago Palm (*Metroxylon sagu* Rottb.)

Lim Whye Kit Leonard

A thesis submitted

In fulfillment of the requirements for the degree of Doctor of Philosophy

(Genetics)

Faculty of Resource Science and Technology

UNIVERSITI MALAYSIA SARAWAK

2022

DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Malaysia Sarawak. Except where due acknowledgements have been made, the work is that of the author alone. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.



.....

Signature

Name: Lim Whye Kit Leonard

Matric No.: 19010029

Faculty of Resource Science and Technology

Universiti Malaysia Sarawak

Date: 7 February 2022

ACKNOWLEDGEMENT

This thesis could not have been come to fruition without the robust assistance and encouragement from many individuals. I would like to take this opportunity to express my heartfelt appreciation to each of them.

First and foremost, I would like to thank God for all the spiritual strength, wisdom and good health granted to me throughout this research.

My deepest gratitude goes to my supervisor, Dr. Chung Hung Hui who have persistently and selflessly mentored me with his overwhelming support and guidance. His never-ending impartment of expertise and knowledge are the keys towards the success of this project. I also would like to thank my co-supervisor, Assoc. Prof. Dr. Hasnain Hussain for his priceless insights and endless encouragement.

Next, I would like to express my exceptional appreciation to my lab mates: Melinda, Mindy, Tiing Tiing, Nabeel, Danial, Joslyn, Pei Ni, Zi Xuan and Nabila, who have voluntarily offered numerous types of aid and inspiration during my hard times. My sincere thank you goes to my friends: John Wong, Delvino, Anarrin, John Chua, Yew, Carlson, Mugu, Simon, Soon Sen and Chia Hong for their motivations.

Last but not least, I am greatly beholden to my dear family members: my wife Teh Zy Ying, my dad Lim Loong Kien, my mum Liew Hong Yen and my sister, Lim Zhi Vern Joanne for their unceasing love. Without them, my journey as a Doctor of Philosophy student would be very much tougher and challenging. Thank you once again to everyone who have helped in making this thesis a reality.

ABSTRACT

The sago palm (*Metroxylon sagu* Rottboll) is a halophytic tree and the ultimate solution to two major global issues: food security and the rapid loss of saline free agricultural lands. This palm can produce high amount of starch, besides being resistant to saline and almost disease free. Unfortunately, this valuable palm is currently underappreciated by the scientific community and the genomic landscape of this palm remained scarce to date. The aim of this study is to sequence and characterise the sago palm chloroplast genome, document the organellar genome copy number changes across various developmental stages, organs, localities and phenotypes as well as sequence the entire nuclear genome of sago palm. The sago palm chloroplast and nuclear genome sizes determined in this study are 157,300 bp and 616,599,600 bp respectively, yielded via next generation sequencing technologies. The organellar genome copy number investigation employing the combinatorial approach of both qPCR and long-PCR assays had unearthed that the mature leaf is generally most enriched with unimpeded plastome and mitogenome among other organs. The *Plawei Mani* growth stage was discovered to have the highest unimpeded organellar genomes copy numbers (over 180,000) among the five major growth stages. The trunking phenotype was proven to be superior to that of the spiny and non-trunking phenotypes in terms of organellar genomes copy numbers (over 188,000). Interestingly, the organellar genome copy numbers across four main Sarawak localities in Malaysia are indifferent statistically. This study serves as a groundwork for preliminary sago palm genomic landscape to support future haplotype and genome wide association studies.

Keywords: *Metroxylon sagu*, chloroplast genome, nuclear genome, organellar genome, genomic landscape

Lanskap Genomik Palma Sagu (*Metroxylon sagu* Rottb.)

ABSTRAK

Palma sagu (Metroxylon sagu Rottboll) ialah pohon halofit dan penyelesaian dua masalah global utama: keselamatan makanan dan kehilangan tanah pertanian bebas garam yang pantas. Pokok palma ini dapat menghasilkan banyak kanji, selain lasak terhadap garam dan hampir bebas penyakit. Malangnya, palma ini kini kurang dihargai oleh komuniti saintifik dan lanskap genomik palma ini masih belum ada. Tujuan kajian ini adalah untuk menjujuk dan mencirikan genom kloroplas sagu, mendokumentasi perubahan bilangan salinan genom organel di pelbagai peringkat perkembangan, organ, kawasan dan fenotip serta menjujuk keseluruhan genom nuklear palma sagu. Saiz genom kloroplas dan nuklear palma sagu yang ditentukan dalam kajian ini ialah 157,300 bp dan 616,599,600 bp, dihasilkan melalui teknologi penjujukan generasi depan. Penyelidikan salinan genom organel menggunakan cara kombinatori ujian qPCR dan PCR panjang telah menemukan bahawa daun matang pada umumnya paling diperkaya dengan plastom dan mitogenom antara organ-organ lain. Tahap Plawei Mani didapati mempunyai salinan genom organel tertinggi antara lima peringkat pertumbuhan utama (melebihi 180,000). Fenotip berbatang terbukti lebih unggul daripada fenotip berduri dan bukan berbatang dari segi salinan genom organel (melebihi 188,000). Menariknya, nombor salinan genom organel di empat kawasan utama Sarawak di Malaysia tidak berbeza secara statistik. Secara keseluruhan, kajian ini berfungsi sebagai landasan dalam membentuk lanskap genomik awal palma sagu untuk menyokong kajian haplotip dan seluruh genom pada masa depan.

Kata kunci: Metroxylon sagu, genom kloroplas, genom nuklear, genom organel, lanskap genomik

TABLE OF CONTENTS

	Page
DECLARATION	i
ACKNOWLEDGEMENT	ii
ABSTRACT	iii
ABSTRAK	iv
TABLE OF CONTENTS	v
LIST OF TABLES	xi
LIST OF FIGURES	xii
LIST OF ABBREVIATIONS	xv
CHAPTER 1: INTRODUCTION	1
1.1 Background	1
1.2 Problem Statement	2
1.3 Objectives	3
1.4 Sago Palm (<i>Metroxylon sagu</i> Rottb.)	4
1.5 Organellar Genomes	5
1.5.1 Chloroplast Genome	7
1.5.2 Mitogenome	8
1.5.3 The Significance of Organellar Genome Copy Numbers on Sago Palm	10
1.6 Research Landscape of Sago Palm	12

1.6.1	Sago Food Manufacturing and Polymer Production	13
1.6.2	Microbiology of Sago Palm	14
1.6.3	Sago Bioprocess Operation	16
1.6.4	Sago Waste Superintendence	18
1.6.5	Sago Palm Pest Control	19
1.6.6	Sago Palm Computational Biology	20
1.6.7	Phenotypic Variation Studies of Sago Palm	21
1.6.8	Sago Palm Population Studies	22
1.6.9	Sago Palm Genetics and Omics	24
1.6.10	The Importance of Sago Palm Genetic Research	26
CHAPTER 2: CHLOROPLAST GENOME SEQUENCING AND		
	CHARACTERIZATION	28
2.1	Introduction	28
2.2	Methodology	30
2.2.1	Genome Sequencing, Assembly and Validation	30
2.2.2	Genome Annotation	31
2.2.3	Microsatellites and Long Repeats Analysis	31
2.2.4	Internal Repeat Expansion and Contraction Analysis	32
2.2.5	Comparative and Divergence Genomic Analysis	32
2.2.6	RNA Editing Analysis	33

2.2.7	Phylogenetic Analysis	33
2.3	Results	33
2.3.1	Genome Assembly and Validation	33
2.3.2	Chloroplast Genome Organisation	35
2.3.3	Analysis of Long and Short Repeats	41
2.3.4	Inverted Repeat Contraction and Expansion	43
2.3.5	Comparative Chloroplast Genome Analysis	45
2.3.6	RNA Editing	54
2.3.7	Phylogenetic Analysis	58
2.4	Discussion	61
2.4.1	Chloroplast Genome Organisation	61
2.4.2	Nucleotide Variation Analysis	61
2.4.3	Phylogenetic Analysis	64
CHAPTER 3: ORGANELLAR GENOME PROFILES ACROSS ORGANS, GROWTH STAGES, PHENOTYPES AND LOCALITIES		65
3.1	Introduction	65
3.2	Methodology	67
3.2.1	Sampling and DNA Extraction	67
3.2.2	Long-PCR	68
3.2.3	Real-time PCR	69

3.2.4	Statistical Analysis	71
3.3	Results	71
3.3.1	Visualization of Long-PCR Bands	71
3.3.2	Organellar Genome Analysis Across Different Organs	72
3.3.3	Organellar Genome Analysis Across Different Growth Stages	74
3.3.4	Organellar Genome Analysis Across Different Phenotypes	75
3.3.5	Organellar Genome Analysis Across Different Localities in Sarawak	77
3.4	Discussion	80
3.4.1	Organellar Genome Analysis Across Different Organs	80
3.4.2	Organellar Genome Analysis Across Different Growth Stages	82
3.4.3	Organellar Genome Analysis Across Different Phenotypes	83
3.4.4	Organellar Genome Analysis Across Different Localities in Sarawak	84
3.4.5	Interspecies Comparison of Organellar Genome Counts	85
CHAPTER 4: GENOME SURVEY ANALYSIS		86
4.1	Introduction	86
4.2	Methodology	87
4.2.1	Sampling and DNA extraction	87
4.2.2	Sequencing and Quality Control	87
4.2.3	K-mer Analysis	88
4.2.4	<i>De novo</i> Genome Assembly and Genome Completeness Assessment	88

4.2.5	Microsatellite Analysis	89
4.2.6	Starch-related Homology Search	89
4.2.7	Identification of rRNA Cistron Components and Phylogenetic Analysis	90
4.3	Results	90
4.3.1	Sequencing Data Statistics	90
4.3.2	K-mer Analysis	91
4.3.3	GC Content and Distribution Pattern	93
4.3.4	<i>De novo</i> Genome Assembly and Genome Comprehensiveness Assessment	94
4.3.5	Small Sequence Repeat (SSR) Analysis	94
4.3.6	Starch-related Homology Analysis	95
4.3.7	Characterisation of rRNA Cistrons and Phylogenetic Analysis	98
4.4	Discussion	100
4.4.1	Genome Length Approximation	100
4.4.2	Genome Heterozygosity, Repeated Sequences Percentage and GC Content	101
4.4.3	Genome Comprehensiveness Assessment	103
4.4.4	Genome-wide Microsatellite Analysis	103
4.4.5	Starch-related Homology Analysis	104
4.4.6	Identification of rRNA Cistron Components and Phylogenetic Analysis	105
	CHAPTER 5: CONCLUSION AND RECOMMENDATIONS	108
5.1	Conclusion	108

5.2	Recommendations	109
-----	-----------------	-----

	REFERENCES	110
--	-------------------	-----

	APPENDICES	157
--	-------------------	-----

LIST OF TABLES

	Page
Table 2.1	Genes encoded by the sago palm chloroplast genome. 36
Table 2.2	The sago palm chloroplast genome features. 37
Table 2.3	The list of genes along with their respective introns in the sago palm chloroplast genome, depicting exon and intron lengths. 38
Table 2.4	Codon usage of both <i>Metroxylon</i> species chloroplast genomes. Ms: <i>M. sagu</i> ; Mw: <i>M. warburgii</i> . 38
Table 2.5	The frequency of discovered microsatellite motifs across the <i>M. sagu</i> and <i>M. warburgii</i> chloroplast genomes. 43
Table 2.6	SNP identified from the comparison between <i>M. warburgii</i> and <i>M. sagu</i> chloroplast genomes. 49
Table 2.7	Indel comparison across both <i>Metroxylon</i> species chloroplast genomes. Ms: <i>M. sagu</i> ; Mw: <i>M. warburgii</i> . 53
Table 2.8	RNA editing sites in sago palm chloroplast transcripts. 55
Table 4.1	List of protein sequences utilised for DIAMOND protein BLASTx with their respective GenBank accession numbers. 89
Table 4.2	The summary of general genomic information of sago palm nuclear genome. 91
Table 4.3	Leading similarity scoring scaffolds (cut-off value 70%) with their aligned length greater than 100 aa, along with their respective GenBank accession number, E-value, identity percentage and amino acid length. 97
Table 4.4	Leading nucleotide hits of sago palm rRNA cistrons, alongside their similarity percentage, E-value and amino acid length. 98

LIST OF FIGURES

	Page
Figure 2.1 The circular gene map of the sago palm chloroplast genome. The genes found on the internal of the bigger circle are transcribed clockwise whereas the genes located on the external of the bigger circle are transcribed anticlockwise. The GC content was illustrated by the darker grey region whereas the AT/U content was depicted by the lighter grey region. LSC, large single copy region; IR, inverted repeat region; SSC, small single copy region.	34
Figure 2.2 The amino acid frequencies in protein-coding genes of both <i>M. sagu</i> and <i>M. warburgii</i> chloroplast genomes.	40
Figure 2.3 The distribution of SSRs in the chloroplast genomes of <i>Metroxylon sagu</i> and <i>Metroxylon warburgii</i> . (A) Number of varying SSR types in both <i>Metroxylon</i> species chloroplast genomes; (B) The distribution pattern of SSRs across different chloroplast genomic regions in both <i>Metroxylon</i> species genomes; (C) The distribution pattern of SSRs in coding and non-coding regions of the two <i>Metroxylon</i> species chloroplast genomes.	42
Figure 2.4 Long repeat sequences analysis on <i>M. sagu</i> and <i>M. warburgii</i> chloroplast genomes. (A) Long repeats types and frequency; (B) Long repeats frequency by length.	43
Figure 2.5 Comparison of IR/LSC and IR/SSC border positions across five Arecaceae family members chloroplast genomes. JLB: LSC/IRb junction; JSB: IRb/SSC junction; JSA: SSC/IRa junction; JLA: IRa/LSC junction.	44
Figure 2.6 Comparison of chloroplast genomes across the five Arecaceae family members, with <i>M. sagu</i> as reference, employing mVISTA alignment program. Dark grey arrows represent gene orientations, dark grey bars indicate exons, light grey bars depict non-coding sequences and white spaces represent differences across the chloroplast genomes.	46
Figure 2.7 Sliding window analysis of the complete chloroplast genomes. Window length: 600 bp; Step size: 200 bp. X-axis: midpoint positions of a window; Y-axis: nucleotide diversity of each window. (A) Nucleotide diversity across <i>M. warburgii</i> and <i>M. sagu</i> . (B) Nucleotide diversity across <i>E. tristis</i> , both <i>Metroxylon</i> species, <i>P. elata</i> and <i>M. flexuosa</i> .	48
Figure 2.8 The distribution of RNA editing sites across sago palm chloroplast genes.	58

- Figure 2.9 The maximum likelihood phylogenetic tree of Arecaceae family members inferred from chloroplast protein-coding genes, with two outgroup gymnosperms (*Typha latifolia* and *Hanguana malayana*) included into analysis. The labels on the right represent the subfamilies of the Arecaceae family whereas the labels on the left represent Calamoid and Coryphoid tribes, with their respective genome sizes respectively. 60
- Figure 3.1 Long-PCR amplicons of mitogenome and plastome DNA across a variety of organs, growth stages and phenotypes of sago palm. Pt: plastome DNA; Mt: mitogenome DNA; ML: mature leaf; NT: non-trunking; S: spiny; YS: young shoot; P: pneumatophore; AL: aged leaf; YR: young root; G1: *Plawei* stage; G2: *Plawei Mani* stage; G3: *Bubul* stage; G4: *Angau Muda* stage; G5: *Angau Tua* stage. 72
- Figure 3.2 (A) The organellar genome copy number across different organs of sago palm established via long-PCR and qPCR. Statistical analysis was implemented using one-way ANOVA and post-hoc Dunn's Test with $p < 0.05$. Significantly distinct data are denoted with different alphabets and the fold differences to that of the nuclear genome copy number are depicted beside the alphabets. (B) The long-PCR/qPCR ratio graph displaying the plastome and mitogenome long-PCR/qPCR ratio across different organs of sago palm. mt: mitogenome; pt: plastome. 73
- Figure 3.3 (A) The organellar genome copy number across different growth stages of sago palm established via long-PCR and qPCR. Statistical analysis was implemented using one-way ANOVA and post-hoc Dunn's Test with $p < 0.05$. Significantly distinct data are denoted with different alphabets and the fold differences to that of the nuclear genome copy number are depicted beside the alphabets. (B) The long-PCR/qPCR ratio graph displaying the plastome and mitogenome long-PCR/qPCR ratio across different growth stages of sago palm. mt: mitogenome; pt: plastome. 75
- Figure 3.4 (A) The organellar genome copy number across different phenotypes of sago palm established via long-PCR and qPCR. Statistical analysis was implemented using one-way ANOVA and post-hoc Dunn's Test with $p < 0.05$. Significantly distinct data are denoted with different alphabets and the fold differences to that of the nuclear genome copy number are depicted beside the alphabets. (B) The long-PCR/qPCR ratio graph displaying the plastome and mitogenome long-PCR/qPCR ratio across different phenotypes of sago palm. mt: mitogenome; pt: plastome. 77
- Figure 3.5 (A) The organellar genome copy number across different main localities of sago palm in Sarawak, Malaysia established via long-PCR and qPCR. Statistical analysis was implemented using one-way ANOVA and post-hoc Dunn's Test with $p < 0.05$. Significantly

different data are conveyed with different alphabets and the fold differences to that of the nuclear genome copy number are indicated beside the alphabets. Non-trunking phenotype samples were not harvested from Pusa and Samarahan due to unobtainability. (B) The long-PCR/qPCR ratio graph illustrating the mitogenome and plastome long-PCR/qPCR ratio across different main localities of sago palm in Sarawak, Malaysia. Non-trunking phenotype samples were not harvested from Pusa and Samarahan due to unobtainability. mt: mitogenome; pt: plastome. Bars denote Sarawak localities.

79

Figure 4.1 The distribution curves of K-mer. (A) K21: K-mer value 21; (B) K25: K-mer value 25; K31: (C) K-mer value 31. GenomeScope parameters: read length = 150 bp; max K-mer coverage = 1000.

92

Figure 4.2 The GC depth graph of sago palm. X-axis: GC percentage; Y-axis: sequencing depth.

93

Figure 4.3 The constitution of dinucleotide, trinucleotide, tetranucleotide and pentanucleotide SSRs identified from the sago palm genome paired-end sequencing reads.

95

Figure 4.4 The SSR percentage bar graphs with the chosen top six frequency SSRs from each group: (A) dinucleotide repeats, (B) trinucleotide repeats, (C) tetranucleotide repeats and (D) pentanucleotide repeats.

95

Figure 4.5 The genome-wide starch-related protein hits predicted utilising the sago palm genome scaffolds.

97

Figure 4.6 The ITS2 maximum likelihood phylogenetic tree of 120 palms (containing sago palm) with Kimura 2 + G model and 1000 bootstrap repetitions.

99

LIST OF ABBREVIATIONS

CTAB	Cetyl Trimethylammonium Bromide
DNA	Deoxyribonucleic Acid
EDTA	Ethylenediaminetetraacetic Acid
IR	Inverted Repeat
ITS	Internal Transcribed Spacer
LSC	Large Single-Copy Region
PCR	Polymerase Chain Reaction
RNA	Ribonucleic Acid
SNP	Single Nucleotide Polymorphism
SSC	Small Single-Copy Region
SSR	Short Sequence Repeat
TAE	Tris Acetate EDTA

CHAPTER 1

INTRODUCTION

1.1 Background

The sago palm (*Metroxylon sagu* Rottboll) is one of the most underutilised and underrated eco-economic food crop of the future. It has several superior characteristics that are unique and advantageous as compared to other commercial and conventional food crops. To name a few, its high starch yield as well as its hardy characteristics like salt resistance and disease tolerance (Karim et al., 2008; Ehara, 2009). Its starch is completely protected within the woody trunk of the tree and it is less vulnerable to pest manifestation. Previously, many researchers have been focusing on studying its starch quality, salt resistance, pest resistance as well as disease resistance physiologically (Schuiling & Flach, 1985; Haryanto & Pangloli, 1992; Notohadiprawiro & Louhenapessy, 1992; Djoeftic, 1999; Purwani et al., 2006; Agus & Subiksa, 2008; Bintoro et al., 2010; Okazaki & Sasaki, 2018; Purwoko et al., 2019; Roslan et al., 2020) but the focus on the genomic aspect is still lacking to date.

The genetic and genomic aspects of plants have been proven powerful in elucidating various underlying functional pathways and mechanisms (Li et al., 2006; Kumar et al., 2014; Ma & Li, 2015). The success of employing both organellar and nuclear genome data in enhancing our comprehension and elucidation of the essential crop metabolic pathways was observed in various key crops like rice, corn, cassava, wheat and sorghum. Distinctive starch bioproduction pathways of cassava were unearthed across varying root developmental stages by Saithong et al. (2013) with the availability of the

entire genomic sequences. Campbell et al. (2016) had successfully identified the effects of domestication on the starch biosynthesis pathway based on whole sorghum genome data. The identification of the contributors towards the starch content regulation pathway in corn kernels was also succeeded through the genome-wide association study (Liu et al., 2016). The detection of the differential starch accumulation rate in endosperm of different rice cultivars, *japonica* and *indica*, was also made possible with the presence of whole genome data (Inukai, 2017). The genome-wide approach had also magnificently assisted in the discovery of key transcription factors orchestrating the reserve starch biosynthesis in wheat (Gu et al., 2021).

1.2 Problem Statement

The sago palm is one imperative player in various fields of industry, contributing significantly towards food security and economy in the past, present and future. However, there is several gaps in research on this cash crop that remained earthed to date and this had impeded the progress to establish this food crop as one of the most promising commodities to human race. Two significant aspects that requires dire attention are none other than the genetic and genomic aspects of the sago palm whereby there is only a handful of research tapping on only the tip of the iceberg.

Interestingly, the sago palm starch yield was previously found to be three to four times more than that of corn, wheat and rice as well as 17 times more than that of the cassava (Karim et al., 2008). Adding to its advantage is that the sago palm is highly tolerable to brackish environment unlike all the aforementioned conventional food crops. The research question is whether the genomic landscape of sago palm could provide us

with more information on the mystery behind its high starch yield coupled with high salt and disease resistance. It is hypothesised that understanding the basic food-producing and defence capability of the sago palm would lead to the comprehension of the contributors orchestrating the metabolic pathways that caused the large discrepancy in starch yield, salt tolerance and disease resistance. This is where the genetic and genomic aspects of the sago palm come into picture as indispensable puzzle pieces in deciphering its unknown starch biosynthesis and salt resistance pathways that make it superior of other food crops.

1.3 Objectives

The genetic and genomic aspects of the sago palm are huge knowledge gaps waiting to be filled urgently. To answer the research question above, the hypothesis of this research was centered around deciphering the genomic landscape of the sago palm via sequencing and characterization approaches. The scope of this study covers various genomic aspects of the sago palm. Therefore, the objectives of this present study are:

- i. To sequence and characterise the entire chloroplast genome of the sago palm;
- ii. To elucidate the organellar genomes copy number at different organs, growth stages, localities and phenotypes of the sago palm; and
- iii. To conduct genome survey on the entire nuclear genome of the sago palm.

1.4 Sago Palm (*Metroxylon sagu* Rottb.)

The sago palm (*Metroxylon sagu* Rottb.) is a true palm grouped in the order Arecales, family Palmae and subfamily Calamideae (Johnson, 1977). Native to Southeast Asia including Indonesia, Papua New Guinea and Malaysia, this palm can thrive well in clumps and in pure stands, in low-land freshwater swampy areas and tropical rain forests (Johnson, 1977; Husaini et al., 2016). It is deemed as one of the potential underutilised food palms that stores starch right within the stem of palm, unlike other crops that normally store starch in the form of cereals, legumes or tubers (Husaini et al., 2016). Its hardy characteristics are probably one of the reasons that explains for its high starch yield per unit: three to four-fold higher than that of corn, rice and wheat and astonishingly 17 times greater than that of cassava (Karim et al., 2008; Ehara, 2009).

The sago palm can grow up to the height of 15 metres and 75 centimetres thickness topped with a massive crown, trunk formation is only initiated at the third to fourth year of palm growth (Kiew, 1977; Kueh, 1977). Within their vegetative phase of seven to fifteen years, photosynthetically produced nutrients from the leaves are channelled to the trunk for storage in the form of starch which upon palm maturity, saturates the trunk up till the crown (Lim, 1991). Flowering occurs at maturity (15 years) and is followed by fruit formation, the falling of mature fruits off the palm is an early indication of palm death (Kueh et al., 1987). In Sarawak, five growth stages have been characterised on mature sago palm, namely “Plawei” (palm at maximum vegetative growth), “Plawei Manit” (emergence of inflorescence), “Bubul” (development of inflorescence), “Angau Muda” (flowering) and “Angau Tua” (fruiting) (Lim, 1991). The “Angau Muda” stage scores the greatest starch content per trunk (39% to 41% on dry weight basis) (Lim, 1991; Pei-Lang et al., 2006) among all other stages, yielding 216 to 219 kg of starch per dry weight (Hamanishi et al.,

1999) whereas the other three stages before the “Angau Tua” stage had showed no significant differences in terms of starch yield (Lim, 1991).

The sago palm is a palm with high economic value where different parts such as the pith, sap, leaf petiole, frond and even the “hampas” (fibrous residue) can be harvested for valuable uses (Singhal et al., 2008). The living tree itself is considered as a countermeasure to the greenhouse effect (Quat Ng, 2007). The sago starch had been placed into the limelight of the food, textile, polymer and pharmaceutical industries with promising potentials (Radley, 1976; Nuttanan et al., 1995; Ishiaku et al., 2002; Purwani et al., 2006). Besides, the sago fronds are excellent materials for papermaking and thatching with amazing durability (Jamaludin et al., 1995). The sago “hampas” is nothing close to what its name suggests because of its wide potential ranging from substrate for microbial conversion via solid state fermentation (Kumaran et al., 1997) to its role as biosorbent (Vickineswary et al., 1994) and activated carbon for industrial wastewater treatment (Kardirvelu et al., 2004; Vennilamani et al., 2005). With more functions of this palm being added to the list at an unprecedented pace, the genetic makeup of this palm might be the key towards better understanding of the driver behind these superb phenotypes of the sago palm.

1.5 Organellar Genomes

The organellar genomes, namely the mitogenome and plastome, are indispensable in various biological and developmental processes. These genomes are exposed to damage caused by various reactive by-products such as reactive oxygen species (ROS) and UV light, because of the metabolism pathways they are involved in (Boecs et al., 2010;

Alexeyev et al., 2013; Kumar et al., 2014). Unlike their nuclear counterpart, their highly polyploid nature favour for the degradation of damaged molecules instead of the DNA repair pathway as seen in the nuclear genome (Liu & Demple, 2010). This accounts for their varied copy numbers in terms of mitochondrial DNA and plastid DNA across various developmental stages of plant.

The high copy number of organellar DNA is a strong indication of great demand for organellar ribosomes and the possible way to fulfil this demand is no other than the elevation of rRNA level from the event of genome multiplication (Bendich, 1987). The copy number variations in plants are one of the major drivers in phenotypic variations, orchestrating heritable dissimilar traits between individuals (Zhang et al., 2009; Iskow et al., 2012). These traits enable plants to possess efficient maintenance pathways essential for the battle against oxidative pressure, environmental damage and replication defaults (Gualberto et al., 2014). Along the process, when not lethal to host plant, mitogenome rearrangements takes place, which can lead to severe phenotypes or induce cytoplasmic male sterility (CMS) (Gualberto et al., 2014).

Energy metabolism impacts the mitochondrial and chloroplast genome greater than the nuclear genome. Both fully functional organellar DNA and nuclear DNA are required for photosynthesis and respiration processes at early stages of organellar development (Kumar et al., 2014; Lim et al., 2020). The high copy of organellar DNA deviates from that of the stable nuclear copy, depicting a decrement trend as leaves start to develop and physiological roles of organellar DNA changes (Kumar et al., 2014). Fully functional plastome is preserved to contribute to chloroplast development until the greening of plant where its degradation is facilitated by ROS. Similarly, the fully functional mitogenome is maintained up till the photosynthesis stage before it is not required anymore in mature

green cells (Kumar et al., 2014). In general, the relinquishment of organellar DNA from the plastome is drastic in maize and gradual in other dicot plants such as tobacco, pea and barrel medick (Oldenburg et al., 2006; Shaver et al., 2006; Rowan & Bendich, 2009).

1.5.1 Chloroplast Genome

The chloroplast is one important metabolic powerhouse that generates carbohydrates from solar energy via the process known as photosynthesis. This organelle also plays indispensable roles in essential plant development and physiology processes, namely in the synthesis of phytohormones, fatty acids, amino acids, nucleotides, vitamins and a myriad of other metabolites as well as in the assimilation of nitrogen and sulphur (Daniell et al., 2016; Lim et al., 2020). These metabolites are utilised greatly in aiding the host plant to interact with its environment, for instances, in response to salt, heat, drought, light and viral invasions (Daniell et al., 2016; Lim et al., 2020).

The first chloroplast genome that was sequenced was that of the tobacco (*Nicotiana tabacum*), which was completed in the year 1986 (Shinozaki et al., 1986). Since then, the number of chloroplast genomes available in the public gene database had elevated exponentially with the aid of the advancement of high-throughput sequencing technologies. The chloroplast genome is one massive reservoir with the complete collection of essential genes involved in plant biology and diversity. Furthermore, it provides the resolution in plant species identification and evolutionary relationship. It is also one important puzzle piece to the comprehension of adaptation mechanism of the host plant towards threats like climate and salinity stress, especially in crops. The chloroplast genome is also made interesting with the study of gene transfer events across the other two genomes, mitogenome and nuclear genome.

The potential of the chloroplast genome in facilitating various downstream research are limitless, especially in terms of biotechnology applications such as investigating stress tolerance (Jin et al., 2015), enzyme production (Zhang et al., 2015), nutrition enhancement (Shintani & DellaPenna, 1998), biopharmaceuticals (Apel & Bock, 2009) as well as vaccines against infectious diseases (Daniell et al., 2016). Stress tolerance researches have advanced since the introduction of numerous transgenes to confer insect resistance in a heterogenous of food crops like corn, wheat and soybean; and in the present days, new strategies like double stranded RNAs and RNA interference emerged as strong and interesting alternatives (Jin et al., 2015; Zhang et al., 2015). The utilization of enzymes isolated from plant chloroplast has proven to be one successful story in the fermentable sugar production research line with raw material like lignocellulosic biomass (Verma et al., 2010). The improvement of nutrition in food crops has seen success in crops like tomato and soybean with the increment of vitamin A and E respectively (Shintani & DellaPenna, 1998; Apel & Bock, 2009; Jin et al., 2014). Oral drugs developed from plant materials via biopharmaceutical approach is one major milestone in human clinical studies as it is one valuable solution to all long-standing problems in drug transportation across nations such as cold storage and short shelf-life of drugs (Su et al., 2015). Moreover, chloroplast-derived vaccines are proven to be effective in curing infectious diseases such as malaria and cholera (Davoodi-Semiromi et al., 2010).

1.5.2 Mitogenome

The mitochondrion is the energy factory that supports life via electron transfer chain. This organelle is believed to have prokaryotic endosymbiont origin and is inherited uniparentally (maternally in most land plants). The evolution of plant mitochondrial DNA

is slower than that of their animal counterparts (Dufay et al., 2007), with lower genetic substitution rates, however the structural variations is much higher in plants (Wolfe et al., 1987). The homologous recombination event of the plant mitochondria is one major driving factor of the mitogenome rearrangements, enabling them to evolve in structure by recombination instead of sequence evolution. These reshufflings resulted in intraspecific dissimilarities and one economic essential trait is none other than the cytoplasmic male sterility (CMS) (viable pollen synthesis incapability inherited maternally) in which CMS has been vital in hybrid seeds synthesis in crop breeding (Dufay et al., 2007).

The mitogenomes of land plants are the largest mitogenomes discovered thus far across that of animals and other plants, the first plant mitogenome sequenced is that of liverwort *Marchantia polymorpha* (Oda et al., 1992). Typical mitogenomes of angiosperm (200 to 700 kb, with the largest found 11.3 Mb in *Silene conica* (Sloan et al., 2013) and algae (13 to 96 kb) are at most 100 times larger than that of animals (15-17 kb) (Gualberto & Newton, 2017). One signature of plant mitogenome is its high number of repeated sequences of varying number and sizes (Alverson et al., 2011). Different sizes of repeats have been categorised and characterised for varying functional roles, namely small repeats (< 50 bp) are responsible for illegitimate microhomology-facilitated recombination, intermediate-sized repeats (50 to 500 bp) are pivotal to sporadic ectopic homologous recombination, large repeats (> 500 bp) play roles in recurrent homologous recombination (Arrieta-Montiel et al., 2009; Davila et al., 2011; Gualberto et al., 2014).

The mitogenome DNA copy number per cell in plants (50 to 120 copies per cell in *Arabidopsis* (Preuten et al., 2010)) are generally much lower than that of chloroplast genomic DNA (1 to 1.7 k per mesophyll cell in *Arabidopsis* (Zoschke et al., 2007) as well as their animal accompaniments (around 150 k mitogenome copies in mature oocytes (Wai

et al., 2008)). Interestingly, the mitogenome DNA copy number was unravelled to be rather diminished in comparison to the total mitochondrial count per cell in most *Arabidopsis* samples examined (Preuten et al., 2010). The most obvious mitogenomic inadequacy was discovered in egg cell of females where the mitogenomic copy number mirrored that of the mesophyll cells (60 copies per cell) (Wang et al., 2010). The possible explanation for this phenomenon is that the main mitogenomic DNA is usually partitioned into subgenomes and it is still uncertain that each nucleoid observed encompass a full set of genomes (Preuten et al., 2010), that is, provided that the interconvertible subgenomes remain catenated during replication (Gualberto et al., 2017). The fission and fusion of mitochondria are the basis genome homogenization, where the mitogenome fusion event facilitates the assortment of genetic materials and in turn aid in tentaculate perinuclear mitochondrial assembly during cell division, especially in leaf primordium meristematic cells and shoot apical meristem (Segui-Simarro et al., 2008). In short, the homologous recombination, reshufflings as well as the copy number of the mitogenome made up the major pillars of studies on plant mitogenomes.

1.5.3 The Significance of Organellar Genome Copy Numbers on Sago Palm

One of the major challenges in the manufacturing of sago starch is the maintenance of the starch quality characteristics between batches. The functional and physiochemical characteristics of sago starch are found to have varied across different growth conditions (types of soil), maturity as well as the parts of trunk where starch is extracted from (Karim et al., 2008). For instance, amylose content was found to be lower in upper part of the trunk and at younger stages before the “Angau Tua” stage (Hamanishi et al., 2000; Tie, 2004). Sago starch harvested from palm grown in mineral soil depicted smaller granules and

fewer amylase activity than those harvested from palm grown in acid sulfate (peat) soil (Nozaki et al., 2004). Besides, the relative crystallinity of starch was greater in upper part of the trunk than the lower part and in 14.5-year-old palm than 9-year-old palm (Hamanishi et al., 1999; Hamanishi et al., 2000). The standards of industrial-grade sago starch (MS468) and edible-grade sago starch (MS470) have been set by the Standards and Industrial Research Institute of Malaysia (SIRIM) in 1976 and 1992 respectively. However, the underlying genetic mechanisms contributing to the varying sago starch quality associated with the metabolic storage pathways remains unknown and underexplored.

The growth of plant is motivated by the energy produced by the two organelles in plant cells, namely chloroplasts and mitochondria. The organellar genome play major roles in the development and growth of plant organ in the way that tight regulations and correct functioning are vital at every developmental stage in plant (van Dingenan et al., 2016). The organellar genome copy number changes frequently across various developmental stages in order to cater to the current needs of the plant at the specific developmental timeframe. In maize, the organellar DNA copy number can fluctuate up to 20-fold under the influence of different light exposures at different tissues as the copy number becomes the limiting factor for the expression of some chloroplast genes (Oldenburg et al., 2006; Zheng et al., 2011; Udy et al, 2012). This is one of the many examples indicating the effect of mitogenome stoichiometric variations beside their impact on cytoplasmic male sterility (Li et al., 1988; Arrieta-Montiel et al., 2001), leaf shape variations (Li et al., 1988) as well as plant growth speed (Ma & Li, 2015). Similarly, in this study, it is interesting to tap into the organellar genome copy number in sago palm to further understand how the starch content and quality are related to the copy number.

The identification of the number of organellar DNA copy number at each different growth stages of the sago palm is essential as it allows for the detailed overview of the starch metabolism in this unique plant where starch is stored in its stem. It is still unknown whether the organellar DNA copy number will change at different developmental stages of sago palm as photosynthetic starch are actively being stored in the trunk, a mechanism that is truly one-of-its-kind among other starch producing crops. Furthermore, there have been research focusing on proteomic comparison between trunking and non-trunking sago palm where it is useful to differentiate stunted individuals (devoid of starch in the trunk) (Hussain et al., 2012a). Several proteins were found to be expressed differentially (Hussain et al., 2012a) and the discovery of the differences in organellar copy number between trunking and non-trunking sago palm would further improve the understanding of this phenotype and reduce economic loss in the industry in future.

1.6 Research Landscape of Sago Palm

The sago palm has many superior characteristics to be recognised as one of the best versatile hardy food crops to be planted massively in the future to ultimately curb the global food security issues (Ehara, 2009). The major ones are high salt resistance, high starch yield and high disease resistance, these three main properties of sago palm allow it to outshine other commercial food crops available now in the market. Added to its aforementioned advantages are its utilities in various fields of research as described below, which highlighted on its discovered tip-of-an-iceberg potential and many more that are yet to be unearthed.

1.6.1 Sago Food Manufacturing and Polymer Production

The sago starch is one indispensable imperative soul of the palm tree, in fact, it has been the staple food source for numerous rural and endogenous communities of Indonesia, Papua New Guinea and Sarawak, Malaysia (Greenhill, 2006; Ehara, 2009; Bintoro et al., 2018). The sago starch is mainly composed of high quality sugars which can be extracted from both the trunk and fronds of the tree, and these starch can be feasibly utilised in various food forms, namely primary products like porridge, cake, paste and bread as well as secondary products such as noodles, desserts, sohun, dumplings, crackers, vermicelli, yogurt, sago pearls, table sugars and flour (Takahashi & Hirao, 1992; Hirao & Takahashi, 1996; Hirao et al., 2008; Nishimura, 2008; Yamamoto et al., 2008; Shima et al., 2012; Ahmad, 2013; Ahmad, 2015; Bujang, 2015; Awg-Adeni et al., 2018; Gunawan et al., 2018; Hirao et al., 2018).

The examination of sago starch noodle quality was performed by Purwani et al. (2006) under moisturised heat exposure. Apart from the superb retention of high firmness and elasticity, the sago starch noodles were found to exhibit minimal cooking depletion (Purwani et al., 2006). Shima et al. (2012) explored the potential of sago starch as raw material catalyst in homemade yogurt to promote the cultivation of lactic acid microorganisms. The promising outcome achieved has shed light towards the future establishment of a healthy probiotic alternative (Shima et al., 2012). The harvest sources of white (purified) sago glucose are sago starch as well as sago fibrous residues (hampas) (Bujang, 2015; Monib, 2015). Bujang (2015) claimed that the sago glucose represents a healthier alternative for cane sucrose. The glucose content of sago sugar is extremely high where only 6% of its composition is non-glucose, the sweetness experimentation also revealed that its value is only half to that of cane sucrose (Bujang, 2015; Monib, 2015).

Moreover, the aptitude of sago starch in replenishing energy and promoting recovery has been explored by Che Jusoh et al. (2016) and Ghosh et al. (2010). As a result, an 84% improvement in performance has been reported with the soy-protein added sago starch supplementation in contrast to flavour-matched placebo (Ghosh et al., 2010; Che Jusoh et al., 2016). On the other hand, Ahmad (2015) and Awg-Adeni et al. (2018) attempted to bioproduce sugar from the fronds of the sago palm via enzymatic hydrolysis and the process yielded at least 9% glucose and 16% cellobiose. The biosynthesised prebiotic cellobiose is one organic choice of starting material in tandem with bifidobacteria in the dairy produce fermentation process (Ahmad, 2015; Awg-Adeni et al., 2018). Interestingly, Gunawan et al. (2018) further improved the nutrition quality of sago flour with respect to protein and amylose level by at most 0.041 fold and 0.33 fold respectively. The sago food trend is experiencing a reorientation of cynosure towards promoting sago starch as an economical surrogate to conventional carbohydrate produces.

On the spectrum of polymer synthesis science, Muljana et al. (2016) spearheaded the ground-breaking synthesis of thermoplastic starch polymer from refused palm cooking oil and sago starch via transesterification with the presence of solvent entity (compressed carbon dioxide). The resultant polymer is equipped with elevated hydrophobicity and thermal stability superior to the native sago starch. This research revolutionised the plastic synthesis norm and aimed to support the green campaign in great dependence on replenishable staple raw materials and biodegradable products (Muljana et al., 2016).

1.6.2 Microbiology of Sago Palm

The focus of the microbiology research of sago palm revolves around the microbiotas that exist from within the tree as well as its external surrounding habitat

(Toyota, 2018). The former microbiota orchestrates the starch quality whereas the latter microbiota exerts imperative influence on the growth betterment of the palm tree as a whole (Toyota, 2018). The acquirement of nitrogen from the soil is greatly dependent on the capability of soil to release nutrients to the sago palm under unfertilised situation (Toyota, 2018).

The microbiota landscape encompassing nitrogen-fixing microorganisms living within sago palm was drafted by Shrestha et al. (2007) via various isolation assays and verifications. The sago palm organs sampled and investigated were starch, bark, root, trunk and midrib (Shrestha et al., 2006). The sago palm was found to house nitrogen-fixing microorganisms like *Pantoea agglomerans*, *Bacillus megaterium*, *Enterobacter cloacae*, *Klebsiella oxytoca* and *K. pneumoniae* (Shrestha et al., 2006). The mutualistic behaviour and relationship of these isolated nitrogen-fixing species were further examined via their activities tested in several combinations in comparison to their individual effects (Shrestha et al., 2007). As a result, it has been confirmed that they played mutualistic roles in tandem for the maximal outcome than their single culture counterparts (Shrestha et al., 2007).

The seclusion of diazotrophic nitrogen fixers from the sago palm rhizosphere and pith was performed by Shipton et al. (2010). From their study, nitrogen fixers were less prominently discovered in the pith in contrast to the moist starch of sago palm. They postulated that these bacteria may have soil origin and that the fixed nitrogen availability is the driver of pathogenic and toxigenic microbiota growth within the sago starch (Shipton et al., 2010). An environmental examination of sago palm microbiota in Papua New Guinea unravelled that 75% of samples are high in coliforms (Greenhill, 2006). Besides, faecal coliforms like *Aspergillus*, *Acremonium*, *Penicillium* and *Scytalidium* were also found in sago starch processed in a traditional manner (Greenhill, 2006). They recommended that

the investigation of haemolytic metabolites produced by fungi can further enrich the knowledge gapped on the sago haemolytic illness aetiology in Papua New Guinea (Greenhill, 2006).

1.6.3 Sago Bioprocess Operation

The bioprocess operation technology of the sago starch involves the biosynthesis of lactic acid, bioethanol and sugar to date. The first sago starch bioprocess was pioneered by Lee et al. (1986) using immobilised *Zymomonas mobilis* as well as amyloglucosidase via the processes of instantaneous saccharification in tandem with fermentation, yielding bioethanol. The ultimate goal of bioprocess is the minimization of time and money whilst targeting for high quality and amount of yield through numerous parameter optimizations, namely pH, nitrogen level, starch input amount as well as enzyme level (Bujang et al., 1996; Hipolito et al., 2000; Mel et al., 2010; Hipolito et al., 2012; Niyong, 2012; Bujang, 2015).

The lactic acid fermentation of sago starch was conducted by Hipolito et al. (2000), Shibata et al. (2007), Hipolito et al. (2012) and Bujang et al. (2018), employing lactic acid microorganisms like *Enterococcus faecium* and *Lactococcus lactis* IO-1. Shibata et al. (2007) revealed that the sago starch outshined other starches in terms of lactic acid biosynthesis level apart from the comparable productivity level achieved by sago starch to that of the glucose. The recurring batch fermentation method designed by Hipolito et al. (2012) had yielded elevated lactic acid return in contrast to the simple batch and continuous manner with respect to the lactate level. Interestingly, a low-cost substitute (sago frond sugar) was subjected to bioprocess experimentation by Bujang et al. (2018)

and the fermentation process recorded a towering 85% efficiency mirroring that synthesised via Standard Medium.

A coupled pre-synthesis phase encompassing the fermentation and hydrolysis was established by Mel et al. (2010) to improve the bioethanol yield, employing α -amylase and glucoamylase for liquefaction and saccharification processes accordingly. The parameters optimised in their study for the hydrolysis stage are saccharification period, starch level as well as glucoamylase level whereas the optimal parameters determined at the fermentation stage are inoculum level, agitation and pH (Mel et al., 2010). Niyong (2012) formulated the Ishizaki method for the incessant fermentation of sago starch into bioethanol. Optimal nitrogen levels were recorded for different raw materials like Mieki, corn steep liquor as well as yeast extract (Niyong, 2012). In a feasibility test consisting of various *Clostridium* species that are solvent tolerant by Johnravindar et al., (2017), the NSW strain transcended its other counterparts and this strain is deemed a valuable substitute for *Bacillus coagulans*.

The innovative way of obtaining sugars unswervingly while skipping the intermediate processes of bioethanol and lactic acid synthesis, was invented by Bujang (2015) utilizing raw materials like sago fibrous residues and sago starch. The hydrolysis of sago starch emulating the parameters optimised by Bujang et al. (2000a) was conducted employing thermal stable Termamyl-12OL isolated from *Bacillus licheniformis* at the liquefaction stage while dextrozymes (glucoamylase and pullulanase combination) were exploited for saccharification process (Bujang et al., 2000b; Bujang & Jobli, 2002; Bujang, 2015). The glucose recovery rate for the optimal sago starch input determined previously (Bujang et al., 2000a; Bujang et al., 2004) was impressively achieved at 100%, provided that pH 4.5 and pH 6.5 were maintained throughout the saccharification and liquefaction stages respectively (Bujang, 2015). The bioprocess technology of sago starch is shifting

towards the utilization of reusable immobilised enzymes to further improve product yield and minimise the synthesis expenses.

1.6.4 Sago Waste Superintendence

There are two states of sago waste: sago liquid wastewater as well as sago fibrous solid waste (also known as the hampas). There are only 30% to 40% non-starch contents in the sago fibrous solid waste (Vickineswary et al., 1994) and the hampas are valuable resources in the manufacturing of biosorbent, fermentable sugars, animal feed, laccase via solid state fermentation as well as mushroom compost (Kumaran et al., 1997; Awg-Adeni et al., 2010; Linggang et al., 2012). On the other hand, the sago wastewater is also essential in the cultivation of *Rhodopseudomonas palustris* and *Spirulina platensis* (Vickineswary et al., 1997; Phang et al., 2000).

In 1997, Vickineswary et al. (1997) utilised 50% sago wastewater as growth medium for strain B1 of *R. palustris* which this bacterium provides nutrition feed for prawn cultivation. On the side note, the COD (chemical oxygen demand) of the cultivation water has been diminished by 77% in total (Vickineswary et al., 1997). In addition, Phang et al. (2000) employed an upflux filled bed digester to farm *Arthrospira* (*S. platensis*) in sago wastewater. Furthermore, the quantity of phosphate and ammoniacal-nitrogen have been lessened by 99.9% and 99.4% respectively along with the decline of COD value by 98% in sum subsequent to digestion (Phang et al., 2000). Besides, it is utmost imperative to implement sago wastewater processing prior to releasing them into the external environment. The fluidised bed reactor was applied by Saravanane et al. (2001) to treat sago wastewater and as a result a 82% maximum efficiency was achieved where the resultant sludge is feasible for agricultural usage. Only a minimal of 30% of the total solid

retained subsequent to the sago wastewater treatment utilizing HUASB (Hybrid Upflow Anaerobic Sludge Blanket) and the COD level has been decreased by up to 91% (Banu et al., 2006). The improved HUASB system with the aid of efficient microbiota, formulated by Priya et al. (2015), had resulted in lesser total suspended solid retention of 23% with 88% COD diminished. It also has been proven that sago wastewater can be translated into biofuel like biogas at a rate of 0.0994 L/day, apart from the 82% COD elimination as well as 92% BOD eradication (Sangeetha & Sivakumar, 2016).

One prominent approach to treat sago hampas via bioconversion into bioethanol was proposed by Awg-Adeni et al. (2013) and this was emulated by Bukhari et al. (2017). A sum of three enzyme digestion phases was utilised and the resultant glucose production and conversion efficiency have been improved significantly which these results provided a good start for the following bioethanol synthesis (Awg-Adeni et al., 2013). The parameters for both enzymatic and acid hydrolysis were optimised by Bukhari et al. (2017) to further improve the bioethanol yield. They concluded that both nitrogen and nutrients are the impellers of yield and conversion rate as acid hydrolysis outperformed enzymatic hydrolysis with a high score of 98% bioethanol yield. Future research should be focused on environmental friendly approaches in reducing unprocessed waste disposal and at the same time utilizing them for bioenergy synthesis with improved yield and reduced costs.

1.6.5 Sago Palm Pest Control

There are 40 insect species that are capable of exerting negative impacts on sago palm growth identified till date (Smith, 2015). The sago palm leaves are habitually subjected to invasion by locusts and crickets like *Monema flavescens* and Psychidae (Kimura, 1979) while the sago palm trunk is regularly suffering from trunk borer larvae of

sago beetles namely *Graphosoma rubrolineatum* as well as *Oryctes rhinoceros* (Kimura, 1979; Jong, 2006). Additionally, young sago palms tend to suffer from white root diseases but fortunately this disease does not threaten the life of the palm (Smith, 2015). The common pests found in sago palm farm in Indonesia and Sarawak state of Malaysia are sago beetles, hispid beetle larvae as well as termites (Smith, 2015).

Furthermore, the vast infestation of alien pests like *Chilades pandava* have been reported across the globe in the proximity of extensive *Cycas* cultivation such as the sago palm cultivation (Donaldson, 2003). In Taiwan, the COII marker has been utilised to reveal the outbreak origin of this pest across 50 distinctive localities (Wu et al., 2010). High haplotype variety was observed and a sum of 29 hyplotypes have been identified, Wu et al. (2010) strongly emphasised that it is imperative to monitor horticultural farms to preserve native species. Future conservation work will be focusing on demographic backgrounds of both native and alien species in order to sustain pest control in a more environmental-friendly mode.

1.6.6 Sago Palm Computational Biology

The computational biology is a novel field of study to be employed on the sago palm and the potential is huge. To date, there is only single publication found and the power of computing in assisting sago palm related research is enormous when being utilised at its full capabilities. A semi-automated sorting program has been established by Hidayat et al. (2018) in hope to map sago palm real time via object-based Pleiades-1A imagery scrutiny. The parameters selected in this study are in terms of textual, arithmetic, spectral and geometric aspects. These aspects are used in machine learning of SVM

(support vector machine) as well as accuracy improvement of this system, the resultant accuracy achieved was 85% (Hidayat et al., 2018).

They deduced that the computational sago palm classifying is still at its infancy and one of the biggest challenges is the high physical appearance similarity shared between sago palm and other palms like oil palm and nipa. Apart from that, the scattered sago palm cultivation with the absence of uniform spatial outline also poses difficulties in the computational prediction of sago palm distribution (Hidayat et al., 2018). It was recommended that future research can set focus on accuracy improvement of the system via additional parameters like atmospheric, segmentation and geometric adjustments (utilizing multi-strip illustrations) (Hidayat et al., 2018).

1.6.7 Phenotypic Variation Studies of Sago Palm

The typical harvest window of the sago palm falls within the range between six to eight years under standard growth situations, nonetheless a number of palms experience no trunk growth unexpectedly beyond the age of a decade to a decade and a half (Hussain et al., 2012a). These non-trunking sago palm trees deprecate economic worth to the sago palm cultivation business. Some researchers highlighted this dire issue and further investigated several aspects like genetic profile, starch particle dimensions, transcriptome as well as proteome in hope to elucidate the orchestrators of this undesired phenotype (Alias, 2012; Hussain et al., 2012a; Edward, 2013; Kamal, 2014).

The trunking and non-trunking sago palms displayed differential protein expression identities as examined by Hussain et al. (2012a) via proteomic assays. Kamal (2014) extended the research by identifying the differential proteins and classifying them into five

distinctive protein families, determined with the aid of 2-D gel electrophoresis and MALDI-TOF/TOF MS. The protein profiled were found to play roles in physical character development ontology (Kamal, 2014). Moreover, the starch features of non-trunking and trunking sago palms were examined by Alias (2012). Interestingly, the trunking and non-trunking sago palms does not differ significantly in terms of starch granule dimensions and starch composition (Alias, 2012). The granule size of the starch isolated from the leaf fell within the range of one to five μm whereas the trunk starch granule dimension ranged between ten to twenty μm (Alias, 2012). On the other hand, the transcriptomic profiles of the trunking and non-trunking sago palms depicted a diverged pattern (Edward, 2013). The non-trunking sago palm research are shifting towards the search for genetic markers to assist in the preliminary screening of sago palm at the seed and seedling stages in order to diminish economic dissipation.

1.6.8 Sago Palm Population Studies

The genetic variations across various populations of sago palm were inspected employing a myriad of genetic markers, namely chloroplast DNA, RAPD (random amplified polymorphism DNA), methylation state of DNA, isozyme profile and AFLP (amplified fragment length polymorphism) markers (Rauwerdink, 1986; Boonsermsuk et al., 1995; Kjær et al., 2004; Abbas et al., 2010; Novero et al., 2012; Abbas et al., 2017).

Upon inspection of sago palms originating from 15 distinguishing populations across three provinces of southern Thailand via isozyme profile analysis, Boonsermsuk et al. (1995) discovered that these palms displayed restricted ancestry as well as mirroring genetic makeup. This evidence is further strengthened by the PAGE (polyacrylamide vertical slab gel electrophoresis) analysis showing zymogram of peroxidase, sorbitol

dehydrogenase, acid phosphate as well as esterase (Boonsermsuk et al., 1995). The subsequent AFLP investigation conducted by Kjær et al. (2004) in Papua New Guinea as well as the RAPD performed by Ehara (2009) across Philippines, Indonesia and Sarawak, Malaysia, had both yielded deductions similar to that of Rauwerdink (1986): all non-spiny and spiny sago palms belonged to the same species across Indonesia, Mindanao, Malaysia and Philippines.

Employing chloroplast gene markers, the sago palm genetic variety in Indonesia was tapped by Abbas et al. (2010). Among the three sampling localities (namely Papua, Kalimantan and Sulawesi), the haplotype variety and diversity were the greatest in the Papua locality, earning its place as the top sago palm germplasm hub in Indonesia (Abbas et al., 2010). A sum of 11 alleles and ten haplotypes was characterised from the 97 trees sampled (Abbas et al., 2010). In the Philippines, Novero et al. (2012) examined the disparity between non-spiny and spiny sago palms with regards to DNA methylation states. The dC (non-methylated cytosine) and 5dmC (methylated cytosine) were utilised as indicators of DNA methylation, these values were measured by HPLC (high performance liquid chromatography). The spiny sago palms differed significantly from their non-spiny counterparts in terms of DNA methylation rate, suggesting that the spine synthesis in sago palm is associated to the age-related epigenetic process (Novero et al., 2012).

The sago palm genetic diversity research was conducted by Abbas et al. (2017) on natural pollinated progenies. Interestingly, the progenies depicted differential morphologies apart from the RAPD genetic dissimilarities, these results are clear evidence of cross-pollination occurred naturally (Abbas et al., 2017). The explanation behind this phenomenon is that the fruits and seeds utilised in that study have distinct genetic backgrounds (Abbas et al., 2017). Undoubtedly, there is a dearth of reproducibility in

RAPD assay, further DNA fingerprinting verification steps such as SCAR (sequence characterised amplified regions) assay are necessitated to elevate the sensitivity and reproducibility levels in detecting polymorphisms undiscovered by other methods (Bhagyawant, 2016).

1.6.9 Sago Palm Genetics and Omics

The molecular biology aspect of the sago palm encompasses the study of genetics, metabolomics, genomics, transcriptomics as well as proteomics. From the genetic point of view, a handful of genes from the sago palm genome were identified and characterised, namely chloroplast *petD*, xylanase, GA20-oxidase, chitinase, ADP-glucose pyrophosphorylase and fructose-1,6-bisphosphate (Jamel et al., 2011; Roslan & Anji, 2011; Wee & Roslan, 2012b; Lim et al., 2022). Looking at the nucleotide records (1465 entries) of the sago palm deposited in the GenBank database to date, about 71.9% (1053) of the entries are nucleotide sequences whereas 28.1% (412) of them are EST (expressed sequence tags) sequences. The sago palm GenBank nucleotide entries are accounted for only 0.028% of the maize entries (5,200,499 entries). The enormous disparity in entry number depicted a huge contrary between both of the abovementioned plants albeit the prodigious aptitude of sago palm to emerge as one of the most feasible food crops for its high yield and salt resistance capabilities (Sasaoka et al., 2014).

Wee and Roslan (2012a) explored the sago palm transcriptomics by building the cDNA library composed of ESTs originated from young leaves. A sum of 49 ESTs were isolated and characterised. The subsequent functional annotation and gene ontology analysis unravelled defence-linked properties interrelated with enzymes such as cytosolic ascorbate peroxidase, cysteine protease as well as class I chitinase, these enzymes are

crucial in shielding young palms from damages induced from unrelenting habitat. Aside from that, these ESTs have stress resistance roles associated with ROS-tackling enzymes like ascorbate peroxidase, glutathione S-transferase as well as alcohol dehydrogenase class III (Wee & Roslan, 2012a). On the side note, they revealed that 5.38% of the ESTs identified are sago palm-specific, opening windows for the research of stress and defence associated proteins.

The metabolomic study of the sago palm was performed using NMR (nuclear magnetic resonance spectroscopy) and GC-MS (gas chromatography-mass spectrometry) by Hussain et al. (2012b). Similar to the research objective by Hussain et al. (2012a) on proteomics, they intend to draft the differential metabolomic profiles across non-trunking and trunking sago palms. The sulfite esters, phosphonates as well as haloalkanes are the ones that distinguish the trunking palms from the non-trunking ones (Hussain et al., 2012b). The proteomic research of the sago palm only encompasses the differential examination of trunking and non-trunking sago palms to date, and this has been discussed under the phenotypic variation studies section. Similarly, the genomic research of the sago palm will be discussed under the genetics and genomics of sago palm towards salinity section. Apart from that, Roslan (2012) is looking into the genome exploration to comprehend the starch synthesis and trunk production mechanisms. The sequencing of the nuclear and mitogenome of sago palm in future would greatly assist in the elucidation of sago palm adaptabilities when facing external environmental stress and striving in undesirable habitat. Besides, enhancer research of sago palm based on genomic data would further enhance the elucidation of stress tolerance and disease resistance pathways at a genome-wide molecular level, putting sago palm on par with other plant models like

Arabidopsis thaliana and *Zea mays*. (Zhu et al., 2015; Weber et al., 2016; Marand et al., 2017; Zicola, 2017; Lim et al., 2018; Lozano et al., 2018).

1.6.10 The Importance of Sago Palm Genetic Research

There are various research fields that have placed focus in utilizing the potential of sago palm to its fullest and one of the most prominent research field is the genetic research. In fact, the genetic research is the fundamental foundation for all other research fields mentioned above for several reasons described below.

The genetic research plays major roles in the food and polymer industries with regard to the elucidation of the underlying metabolic pathways that leads to sago starch quality variation (in terms of crystallinity, prebiotic content, amylose content and granule size) harvested at differing growth stage, growth environment, organ, phenotype and locality (Hamanishi et al., 1999; Hamanishi et al., 2000; Nozaki et al., 2004; Tie, 2004; Karim et al., 2008). In the field of microbiology, genetic research can be very informative on the metagenomics of sago palm and also provides clues to sago haemolytic illness aetiology (Greenhill, 2006; Shrestha et al., 2007; Shipton et al., 2010; Toyota, 2018). Despite the maturity in the bioprocess technology, the unresolved starch biosynthesis pathway of the sago palm can only be deciphered via the genetic approach (Hipolito et al., 2000; Mel et al., 2010; Niyong, 2012; Bujang, 2015; Bujang et al., 2018). The sago waste disposal field requires the genetic research to identify the presence of invasive species in the waste solid and liquid before their discharge into the environment (Awg-Adeni et al., 2013; Bukhari et al., 2017).

The pest control research field necessitates the aid of genetic research in the susceptibility investigation of various sago palm phenotypes towards pests via DNA fingerprinting profiling (Wu et al., 2010; Smith, 2015). The computational biology research field greatly necessitates the complementation of the genetic research discoveries in order to further improve the accuracy of *in silico* predictions via several rounds of scrutinised experimental validations (Hidayat et al., 2018). There is no qualm that the phenotypic variation studies rely greatly on the genetic assays to differentiate sago palm phenotypes down to the molecular level (Hussain et al., 2012a; Kamal, 2014). Furthermore, population studies would not be accurate enough without the intervention of genetic markers to complement morphometric characterizations (Abbas et al., 2010; Novero et al., 2012).

The genetic research on the sago palm is still in its infancy but it is picking its pace recently with the technology advancement in genome sequencing. The genomic landscape of the sago palm would definitely lead to the filling of knowledge gaps in various fields of sago palm research. A good start would be the sequencing of the whole genome of sago palm to unravel the sago palm genomic makeup that is still earthed to date.

CHAPTER 2

CHLOROPLAST GENOME SEQUENCING AND CHARACTERIZATION

2.1 Introduction

The chloroplasts are widely found in eukaryotic algae and plant cells, their genomes are formed via endosymbiotic events (McFadden, 2001; Gray, 2004). These organelles play major roles in a myriad of biochemical processes such as starch bioproduction, photosynthesis and salinity resistance (Vothknecht & Westhoff, 2001; Pfister & Zeeman, 2016). The size of plant chloroplast genome can vary between 120 kb and 210 kb, it can house a sum of 110 to 130 genes (Chumley et al., 2006; Ravi et al., 2007; Lin et al., 2010; Dong et al., 2013; Tian et al., 2018). A SSC (small single copy fragment), a LSC (large single copy fragment) as well as two sandwiching IRs (inverted repeat fragments) form the entire circular quadripartite conformation of the chloroplast genome. The predominant inverted repeat length of plant chloroplast genomes falls in the range between 20 kb and 30 kb (Palmer & Delwiche, 1998).

The chloroplast genome is a rich reservoir of genetic markers valuable for research focusing on phylogeny, taxonomy, population genetics, phylogeography as well as DNA fingerprinting (Song et al., 2019). A great advantage that a chloroplast genome offers is that it evolves slower than its nuclear counterpart, apart from uniparental inheritance characteristics (maternal inheritance in most angiosperms) (Song et al., 2019). The chloroplast is one centralised hub hosting numerous imperative biological pathways like translation, transcription, RNA editing as well as photosynthesis and these processes normally involve the participation of majority of the genes within the genome (Bock, 2000; Ahlert et al., 2003; Schmitz-Linnerweber & Barken, 2007; Uthapaisanwong et al., 2012).

The first chloroplast genome sequenced belonged to that of the liverwort where the sequencing work was completed in 1986 (Ohyama et al., 1986). Thenceforth, the GenBank data base chloroplast genome records have grown in numbers exponentially to over 14,000 to date.

The Arecaceae palm family, where the sago palm was categorised under, umbrellaed over 200 genera and beyond 2,500 species. The GenBank repository for chloroplast genomes now housed a sum of 38 complete sequence records, such as *Elaeis guineensis*, *Phoenix dactylifera*, *Pigafetta elata*, *Metroxylon warburgii*, *Cocos nucifera* and *Eugeissona tristis* (Yang et al., 2010; Uthaipaisanwong et al., 2012; Huang et al., 2013; Barrett et al., 2016). The starch made up the major proportion of yield from the sago palm tree and the starch bioproduction pathway is governed by none other than the chloroplast via specialised amyloplasts (Pfister & Zeeman, 2016) that are capable of stockpiling starch in the trunk. The abundant starch yield of the sago palm provides a dire need to unravel the contributors towards this unique trait via the sequencing of the whole chloroplast genome. With the consolidation of whole genome sequences from its nuclear and mitochondrion counterparts, the chloroplast genome data will complete the comprehension of the genomic landscape of the sago palm (Tian et al., 2018). Hence, in this chapter, the chloroplast genome of sago palm was sequenced and characterised to further decipher the underlying knowledge gaps in its starch utilization.

2.2 Methodology

2.2.1 Genome Sequencing, Assembly and Validation

The chloroplast genomic DNA extraction was performed using 30 g of surface-sterilised sago palm leaves (specimen was identified and verified by our plant expert before harvesting from the living germplasm collection area located in Universiti Malaysia Sarawak East Campus: 1°27'52.3"N, 110°27'04.7"E) employing CP-ISO Chloroplast Isolation Kit (Sigma Aldrich, USA). Three young leaf blades from three different lead fronds (three biological replicates) that were harvested from each sago palm tree (total three palms) during the Plawei Manit growth stage (8 years old) were used. The place of origin of this palm chosen is Pusa, Sarawak, Malaysia. The genomic DNA extracted was subjected to quality screening via agarose gel electrophoresis as well as Qubit 2.0 (Thermo Fisher Scientific, USA) to guarantee high concentration and purity.

A total of 1.0 µg genomic DNA was utilised for library preparation using NEBNext DNA Library Prep Kit (NEB, USA) according to manufacturer's protocol and indices were included into each sample. Random shearing of the genomic DNA into 350 bp was conducted prior to end-polishing, A-tailing and ligating with NEBNext adapter for Illumina sequencing. PCR enrichment was done onto these fragments employing P5 and indexed P7 primers before the purification of fragments using AMPure XP System (Beckman Coulter, USA).

Analysis of the resulted libraries was performed to check for size distribution using 2100 Bioanalyzer (Agilent, USA) and these libraries were quantified using real-time PCR. The high-quality libraries were fed into Illumina HiSeq 4000 PE150 sequencers (Illumina, USA) subsequent to pooling based on its expected data volume as well as effective

concentrations. The assembly of sequence reads was conducted using Norgal *de novo* assembler (Al-Nakeed et al., 2017) and QUASt (Gurevich et al., 2013). These sago palm genome contigs were identified via alignment to the *M. warburgii* reference genome utilizing Sequencher software version 4.10.1 (Gene Codes, 2019).

2.2.2 Genome Annotation

The genome annotation of sago palm chloroplast genome was done using GeSeq online tool (Tillich et al., 2017) under the bacterial/plant genetic code with default parameters. BLAST searches were done against *M. warburgii* NCBI RefSeq data prior to comparative analysis between the two genomes for further manual genome annotation verification. All protein-coding sequences were quality checked in respect of amino acids using MEGA X (Kumar et al., 2018) to remove and correct truncated and premature stop codons for functionalities. Annotation of tRNA was performed using tRNAScan-SE v2.0.3 (Lowe and Chan, 2016) with default settings. Gene homologies was verified employing Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehsia et al., 2016). The structural characteristics of the sago palm chloroplast genome map were visualised and constructed via OGDRAW v1.2 (Lohse et al., 2013). MEGA X (Kumar et al., 2018) was used to determine the codon usage, nucleotide composition as well as relative codon usage (RSCU).

2.2.3 Microsatellites and Long Repeats Analysis

The MIcroSAteLLite (MISA) web tool (Beier et al., 2017) was utilised to identify microsatellites within the sago palm chloroplast genome. The search parameters were set to

harvest perfect mono-, di-, tri-, tetra-, penta- and hexanucleotide motifs with minimum 8, 5, 3, 3 and 3 repeats respectively. The long repeats were discovered using REPuter online software (Kurts et al., 2001) in terms of reverse, palindromic, complement and forward repeat sequences at minimum repeat size of 30 bp with high sequence identity (>90%).

2.2.4 Internal Repeat Expansion and Contraction Analysis

A sum of four junctions (SSC/IRa, IRa/LSC, LSC/IRb and IRb/SSC), IRs (IRa and IRb) as well as the two single copy regions (LSC and SSC) were investigated for inverted repeat expansion and contraction across five palm members, namely *Eugeissona tristis*, *Metroxylon sagu*, *Pigafetta elata*, *Metroxylon warburgii* and *Mauritia flexuosa*. The junction borders and sites were determined via sequence homology utilizing IRscope web tool (Amiryousefi et al., 2018) with default parameters.

2.2.5 Comparative and Divergence Genomic Analysis

The five abovementioned Arecaceae family palms, including the sago palm, were subjected to both comparative and divergence analyses. The mVISTA online software (<http://genome.lbl.gov/vista/mvista/about.shtml>) was chosen for this purpose. The mode used was the Shuffle-LAGAN mode (Frazer et al., 2004) for the comparative and divergence genome analysis. The reference genome used for this analysis was the sago palm chloroplast genome. The DnaSP v5.1 program (Librado & Rozas, 2009) was then employed for the nucleotide variability (Pi) enumeration. The parameters set for this analysis were step size 200 bp and window length 600 bp.

2.2.6 RNA Editing Analysis

RNA editing events of sago palm chloroplast genome were predicted using PREPACT v3.0 software (Lenz et al., 2018) with default search mode and the reference genome used is that of *Cocos nucifera* (NC_022417.1).

2.2.7 Phylogenetic Analysis

All protein-coding genes from 39 palms (and two gymnosperm outgroups *Typha latifolia* and *Hanguana malayana*) were selected as input to the phylogenetic analysis. A model test was first conducted employing clustal w in MEGA X (Kumar et al., 2018) for the selection of the best model under the maximum likelihood criterion. The parameters of GTR + I + G nucleotide substitution model and 1000 bootstrap replications were chosen for the maximum likelihood family tree construction via MEGA X (Kumar et al., 2018).

2.3 Results

2.3.1 Genome Assembly and Validation

The chloroplast genome sequencing of sago palm had yielded a sum of 5,629,340 raw contig sequence reads with 99.32% (5,590,920) clean contig reads, 0.67% (37,620) adapter-related reads, 0.01% (800) N-nucleotide-containing reads as well as 0% (0) low quality read. The effective rate was recorded significant at 99.32% whereas the error rate was recorded miniature at 0.03%. The major sequence read population (91.98%) fell within the Q score range of more than 20, followed by 84.54% residing within the Q score > 30. The screening procedure had excluded low quality (Q score < 5) reads, adapter-containing

reads and N-nucleotide enriched reads. The complete chloroplast genome of sago palm was assembled with reference to that of *M. warburgii* using QUASt (Gurevich et al., 2013) and Norgal de novo assembler (Al-Nakeeb et al., 2017) before the visualisation of the complete sago palm chloroplast genome construct (Figure 2.1).

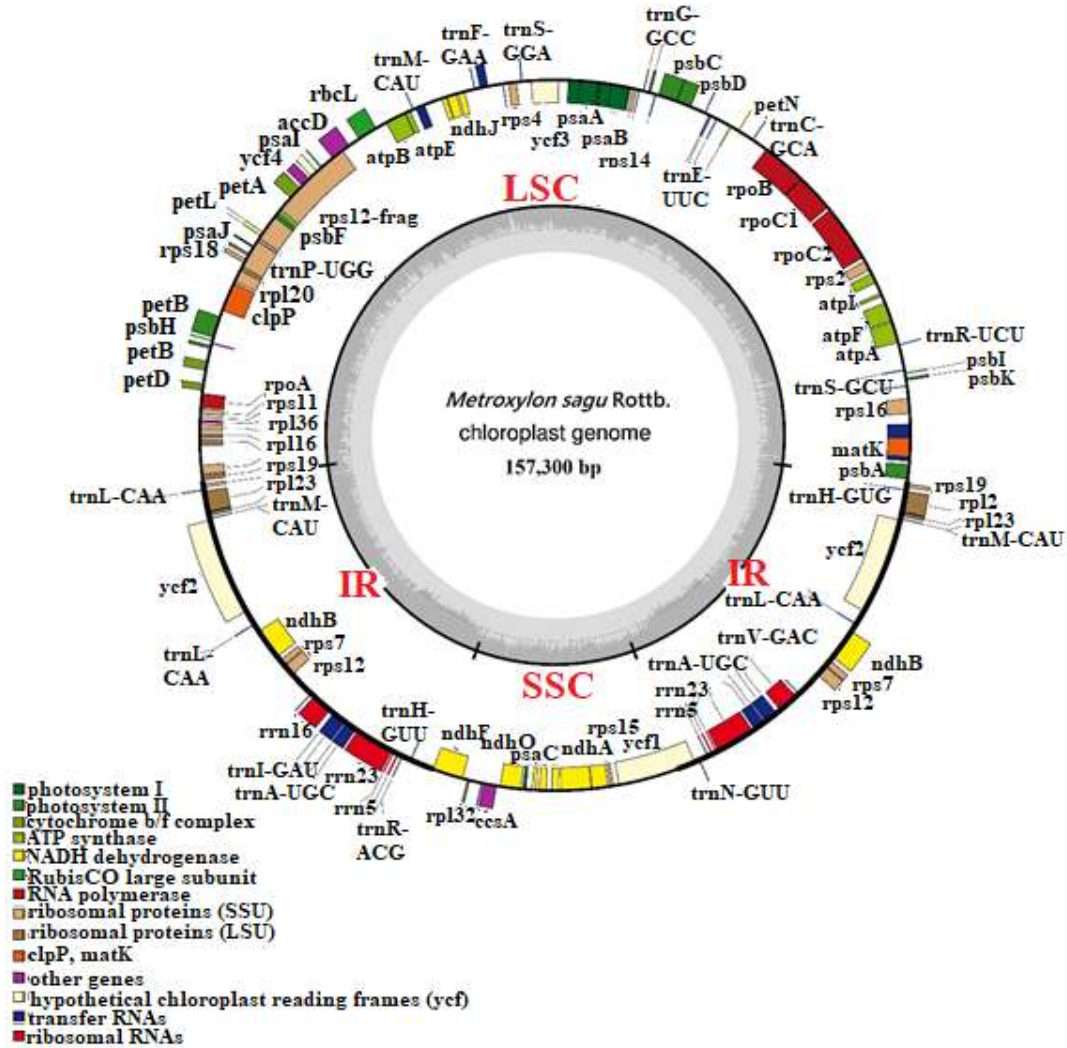


Figure 2.1: The circular gene map of the sago palm chloroplast genome. The genes found on the internal of the bigger circle are transcribed clockwise whereas the genes located on the external of the bigger circle are transcribed anticlockwise. The GC content was illustrated by the darker grey region whereas the AT/U content was depicted by the lighter grey region. LSC, large single copy region; IR, inverted repeat region; SSC, small single copy region.

2.3.2 Chloroplast Genome Organisation

The conformation of the complete sago palm chloroplast genome resembles that of a single circular shape with quadripartite arrangement, encompassing an entire length of 157,300 bp (GenBank accession number: MN309778). The sago palm chloroplast genome consists of 85,257 bp large single-copy region (LSC), two sandwiching 27,245 bp inverted repeats (IRs) as well as 17,533 bp small single-copy region (SSC). There is a total of 113 distinctive genes and 23 duplicative genes residing in the sago palm chloroplast genome. A sum of three pseudogenes (two duplicated in IR), 29 tRNA genes (11 duplicated in IR), four ribosomal RNA genes (all four duplicated in IR) as well as 80 functional genes (six duplicated in IR) were determined (Table 2.1).

Around 52.28% of the sago palm chloroplast genome houses the protein-coding sequences, amounting to 82,233 bp whereas 40.13% of the genome is inhabited by non-coding sequences (63,127 bp). The rRNA and tRNA genes occupy 5.75% (9,050 bp) and 1.84% (2,890 bp) of the entire chloroplast genome of sago respectively (Table 2.2). The overall AT/U content of the sago palm chloroplast genome was found to be mirroring to that of oil palm, which is 62.6% (Uthaipaisanwong et al., 2012) and this value is highly conserved across other Arecaceae family members such as date palm (62.8%) (Yang et al., 2010) and natangura palm (62.7%) (Barrett et al., 2016).

There are 17 sago palm chloroplast genes that have introns within their sequences in which 88.24% (15) of them contain an intron and the remaining two genes (*ycf3* and *clpP*) contain two introns respectively (Table 2.3). An incomplete duplication of 3' end of *ycf1* was observed in the SSC region of the sago palm chloroplast genome, a phenomenon that is not alien in date palm (Yang et al., 2010), natangura palm (Barrett et al., 2016) as well as oil palm (Uthaipaisanwong et al., 2012). The *ycf1* partial duplication fragment is

1,343 bp long in sago palm and natangura palm, but the length is 1,339 bp and 1,345 bp in date palm and oil palm respectively (Yang et al., 2010; Uthaipaisanwong et al., 2012; Barrett et al., 2016). The hypothetical genes *ycf68* and *ycf15* were defined as non-functional after investigation via sequence homology across other palm species owing to the presence of internal stop codons, as seen in oil palm (Uthaipaisanwong et al., 2012).

Table 2.1: Genes encoded by the sago palm chloroplast genome.

Category	Gene Name
1. Photosystem I	<i>psaA, psaB, ycf3, psaI, ycf4, psaJ, psaC</i>
2. Photosystem II	<i>psbA, psbK, psbI, psbM, psbD, psbC, psbZ, psbJ, psbL, psbF, psbE, psbB, psbT, psbN, psbH</i>
3. Cytochrome b/f	<i>petN, petA, petL, petG, petB, petD</i>
4. ATP synthase	<i>atpA, atpF, atpH, atpI, atpE, atpB</i>
5. NADH dehydrogenase	<i>ndhJ, ndhK, ndhC, ndhB, ndhD, ndhE, ndhG, ndhI, ndhA, adhH, ndhF</i>
6. Rubisco	<i>rbcL</i>
7. RNA polymerase	<i>rpoC1, rpoC2, rpoB, rpoA</i>
8. Large subunit ribosomal protein	<i>rpl33, rpl20, rpl36, rpl14, rpl16, rpl22, rpl2*, rpl23*, rpl32</i>
9. Small subunit ribosomal protein	<i>rps16, rps2, rps14, rps4, rps18, rps11, rps8, rps3, rps19*, rps15, rps7*, rps12*</i>
10. Other proteins	<i>matK, accD, cemA, clpP, infA, ccsA, pbf1</i>
11. Proteins of unknown function	<i>ycf2*, ycf1</i>
12. Ribosomal RNAs	<i>rrn4.5*, rrn5*, rrn16*, rrn23*</i>
13. Transfer RNAs	<i>trnA-UGC*, trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnI-M-CAU, trnG-UCC*, trnH-GUG*, trnI-CAU*, trnI-GAU*, trnK-UUU, trnL-CAA*, trnL-UAA*, trnL-UAG, trnM-CAU, trnN-GUU*, trnP-UGG, trnQ-UUG, trnR-ACG*, trnR-UCU, trnS-GCU, trnS-UGA, trnT-GGU, trnV-GAC*, trnV-UAC*, trnW-CCA, trnY-GUA, trnS-GGA, trnT-UGU</i>
14. Pseudogene	<i>ycf15*, ycf1, ycf68*</i>

*:Two gene copies in the IRs; The *rps12* gene is trans-spliced.

The protein-coding genes of sago palm are made up of 27,411 codons. The codon usage of sago palm was skewed towards a high AT/U (69.7%) bias, particularly at the third codon site (Table 2.2). This is one general occurrence found across most land plants (Clegg et al., 1994). The relative synonymous codon usage (RSCU) of *M. sagu* was compared against *M. warburgii* (the only *Metroxylon* genus member with complete chloroplast genome sequenced) (Table 2.4). The TGG and ATG (tryptophan and methionine) are the only ones depicting no bias (RSCU = 1.00). Serine and isoleucine are the most employed codons in both *Metroxylon* species compared with the higher degrees recorded in *M. warburgii* for both codons (Figure 2.2). One common situation discovered among the preferred synonymous tRNA codons (RSCU > 1.00) is that they are all ended with A or U, except for trnR-AGG and trnL-UUG in both *Metroxylon* species.

Table 2.2: The sago palm chloroplast genome features.

Regions	Positions	Length (bp)	T/U (%)	C (%)	A (%)	G (%)	AT/U (%)
Genome		157,300	31.7	19.0	30.9	18.4	62.6
LSC		85,257	31.9	17.7	34.6	15.7	66.5
IRa		27,245	29.0	21.9	28.6	20.5	57.6
SSC		17,553	34.5	16.2	34.1	15.2	68.6
IRb		27,245	29.0	21.9	28.6	20.5	57.6
Protein coding genes		82,233	31.2	17.8	30.8	20.1	62.0
	1st position	27,411	23.4	18.7	31.1	26.8	54.5
	2nd position	27,411	32.4	20.4	29.6	17.6	62.0
	3rd position	27,411	37.8	14.4	31.9	15.9	69.7
tRNA		2,890	24.8	23.7	21.9	29.6	46.7
rRNA		9,050	18.7	23.7	26.0	31.6	44.7

Table 2.3: The list of genes along with their respective introns in the sago palm chloroplast genome, depicting exon and intron lengths.

Gene	Location	Exon I (bp)	Intron I (bp)	Exon II (bp)	Intron II (bp)	Exon III (bp)
atpF	LSC	411	815	141		
clpP	LSC	228	654	292	847	71
ndhA	SSC	539	1127	553		
ndhB	IR	756	704	777		
rpl2	IR	385	662	431		
rpoC1	LSC	1635	736	414		
rps12	LSC/IR	114	0	230	551	26
rps16	LSC	225	858	36		
ycf3	LSC	161	735	228	728	124
trnA-UGC	IR	38	801	35		
trnC-ACA	LSC	56	569	39		
trnG-UCC	LSC	23	713	48		
trnI-GAU	IR	42	937	35		
trnL-UAA	LSC	35	509	50		
trnStop-UUA	LSC	37	2567	37		
trnS-UGA	LSC	38	509	47		
trnV-UAC	LSC	37	588	39		

Table 2.4: Codon usage of both *Metroxylon* species chloroplast genomes. Ms: *M. sagu*; Mw: *M. warburgii*.

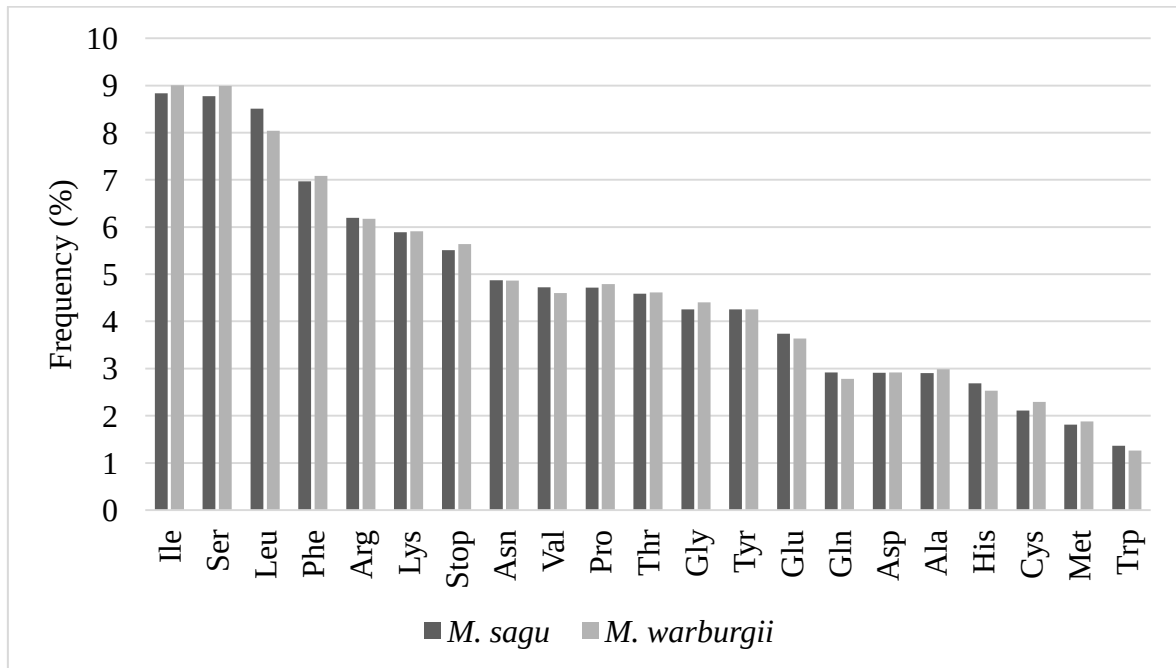
Amino acid	Codon	Count		RSCU		Frequency		tRNA	
		Ms	Mw	Ms	Mw	Ms	Mw	Ms	Mw
Ala	GCA	494	486	1.3	1.24	0.942	0.926	trnA-UGC	trnA-UGC
Ala	GCC	336	349	0.88	0.89	0.641	0.665		
Ala	GCG	229	231	0.6	0.59	0.437	0.440		
Ala	GCU	466	500	1.22	1.28	0.888	0.952		
Cys	UGC	382	436	0.69	0.72	0.728	0.830	trnC-GCA	trnC-GCA
Cys	UGU	726	769	1.31	1.28	1.384	1.465		
Asp	GAC	408	425	0.53	0.55	0.778	0.809	trnD-GUC	trnD-GUC
Asp	GAU	1118	1107	1.47	1.45	2.131	2.108		
Glu	GAA	1338	1353	1.36	1.42	2.551	2.577	trnE-UUC	trnE-UUC
Glu	GAG	623	555	0.64	0.58	1.188	1.057		
Phe	UUC	1511	1532	0.83	0.82	2.880	2.918	trnF-GAA	trnF-GAA
Phe	UUU	2146	2187	1.17	1.18	4.091	4.165		
Gly	GGA	807	835	1.45	1.45	1.538	1.590	trnG-UCC	trnG-UCC

Table 2.4 continued

Gly	GGC	304	324	0.55	0.56	0.580	0.617		
Gly	GGG	492	526	0.88	0.91	0.938	1.002		
Gly	GGU	628	626	1.13	1.08	1.197	1.192		
His	CAC	396	384	0.56	0.58	0.755	0.731	trnH-GUG	trnH-GUG
His	CAU	1015	945	1.44	1.42	1.935	1.800		
Ile	AUA	1577	1657	1.02	1.05	3.006	3.156		
Ile	AUC	1226	1250	0.79	0.79	2.337	2.381	trnI-GAU	trnI-GAU
Ile	AUU	1831	1821	1.19	1.16	3.490	3.468		
Lys	AAA	2067	2089	1.34	1.35	3.940	3.979	trnK-UUU	trnK-UUU
Lys	AAG	1024	1015	0.66	0.65	1.952	1.933		
Leu	CUA	771	703	0.88	0.86	1.470	1.339	trnL-UAG	trnL-UAG
Leu	CUC	680	653	0.78	0.8	1.296	1.244		
Leu	CUG	539	489	0.62	0.6	1.027	0.931		
Leu	CUU	1041	1046	1.19	1.27	1.984	1.992		
Leu	UUA	1105	961	1.27	1.17	2.106	1.830	trnL-UAA	trnL-UAA
Leu	UUG	1097	1072	1.26	1.31	2.091	2.042	trnL-CAA	trnL-CAA
Met	AUG	950	988	1	1	1.811	1.882		
Asn	AAC	802	794	0.63	0.62	1.529	1.512	trnN-GUU	trnN-GUU
Asn	AAU	1752	1760	1.37	1.38	3.340	3.352		
Pro	CCA	799	855	1.29	1.36	1.523	1.628	trnP-UGG	trnP-UGG
Pro	CCC	612	569	0.99	0.9	1.167	1.084		
Pro	CCG	383	403	0.62	0.64	0.730	0.768		
Pro	CCU	681	689	1.1	1.1	1.298	1.312		
Gln	CAA	1047	987	1.37	1.35	1.996	1.880	trnQ-UUG	trnQ-UUG
Gln	CAG	484	474	0.63	0.65	0.923	0.903		
Arg	AGA	1016	1037	1.88	1.92	1.937	1.975	trnR-UCU	trnR-UCU
Arg	AGG	619	611	1.14	1.13	1.180	1.164		
Arg	CGA	615	622	1.14	1.15	1.172	1.185		
Arg	CGC	227	214	0.42	0.4	0.433	0.408		
Arg	CGG	377	355	0.7	0.66	0.719	0.676		
Arg	CGU	396	404	0.73	0.75	0.755	0.769	trnR-ACG	trnR-ACG
Ser	AGC	474	481	0.62	0.61	0.904	0.916	trnS-GCU	trnS-GCU
Ser	AGU	681	695	0.89	0.88	1.298	1.324		
Ser	UCA	873	896	1.14	1.14	1.664	1.707	trnS-UGA	trnS-UGA

Table 2.4 continued

Ser	UCC	863	905	1.13	1.15	1.645	1.724	trnS-GGA	trnS-GGA
Ser	UCG	604	608	0.79	0.77	1.151	1.158		
Ser	UCU	1107	1137	1.44	1.44	2.110	2.166		
Stop	UAA	1091	1151	1.13	1.17	2.080	2.192		
Stop	UAG	774	760	0.8	0.77	1.475	1.447		
Stop	UGA	1026	1051	1.06	1.06	1.956	2.002		
Thr	ACA	705	718	1.17	1.19	1.344	1.367	trnT-UGU	trnT-UGU
Thr	ACC	595	585	0.99	0.97	1.134	1.114	trnT-GGU	trnT-GGU
Thr	ACG	393	380	0.65	0.63	0.749	0.724		
Thr	ACU	713	738	1.19	1.22	1.359	1.406		
Val	GUA	778	750	1.26	1.24	1.483	1.428	trnV-UAC	trnV-UAC
Val	GUC	446	466	0.72	0.77	0.850	0.888	trnV-GAC	trnV-GAC
Val	GUG	433	432	0.7	0.72	0.825	0.823		
Val	GUU	820	767	1.32	1.27	1.563	1.461		
Trp	UGG	716	663	1	1	1.365	1.263	trnW-CCA	trnW-CCA
Tyr	UAC	720	687	0.65	0.62	1.373	1.308	trnY-GUA	trnY-GUA
Tyr	UAU	1510	1547	1.35	1.38	2.878	2.946		

**Figure 2.2:** The amino acid frequencies in protein-coding genes of both *M. sagu* and *M. warburgii* chloroplast genomes.

2.3.3 Analysis of Long and Short Repeats

Short sequence repeats (SSRs) or microsatellites are widely dispersed across the entire chloroplast genome of plants as a single to six nucleotide repeating units (Richard et al., 2008). The SSRs discovered in the sago palm is 193, which is 8 SSRs less than that of its congeneric counterpart, *M. warburgii* (Figure 2.3). The greatest motif population is undoubtedly the mononucleotide motifs, followed by the dinucleotide motifs. The amount of SSRs is highly conserved across both *Metroxylon* species, especially tetranucleotide and trinucleotide motifs. The SSR profile for *M. sagu* is as follow: 164 mononucleotides, 15 dinucleotides, 2 trinucleotides, 10 tetranucleotides, 2 pentanucleotides and one hexanucleotides; whereas the *M. warburgii* depicted the SSR profile: 165 mononucleotides, 18 dinucleotides, 2 trinucleotides, 10 tetranucleotides, 4 pentanucleotides and 2 hexanucleotides. The SSRs of sago palm are mostly located in the non-coding regions and the occurrence rate is more than twice to that of the coding regions. The incidence of SSRs in the LSC is at least four-fold more than other regions like IRs and SSC. On the side note, there is an obvious AT bias observed across all SSRs detected where 78.6% and 80.8% of them are composed of A or T nucleotide in *M. sagu* and *M. warburgii* respectively (Table 2.5).

The long repeats analysis via REPuter (Kurtz et al., 2001) at 90% identity cut-off revealed the presence of only palindromic and forward long repeats in sago palm, amounting to 26 and 24 respectively (Figure 2.4). On the other hand, *M. warburgii* has one additional long repeat type (three reverse long repeats). The amounts of palindromic and forward long repeats of *M. warburgii* are not significantly differed from that of the sago palm, which are 25 and 22, respectively. It is interesting to note that both chloroplast genomes lack complement long repeats. Among the 50 long repeats identified across both

chloroplast genomes, most of them are categorised under the 30-50 bp group. The sago palm has one long repeat less than that of *M. warburgii* within the 30-39 bp size range. The amount of 42-49 bp long repeats is conserved across the *Metroxylon* counterparts whereas their >50 bp long repeats differed by 3.

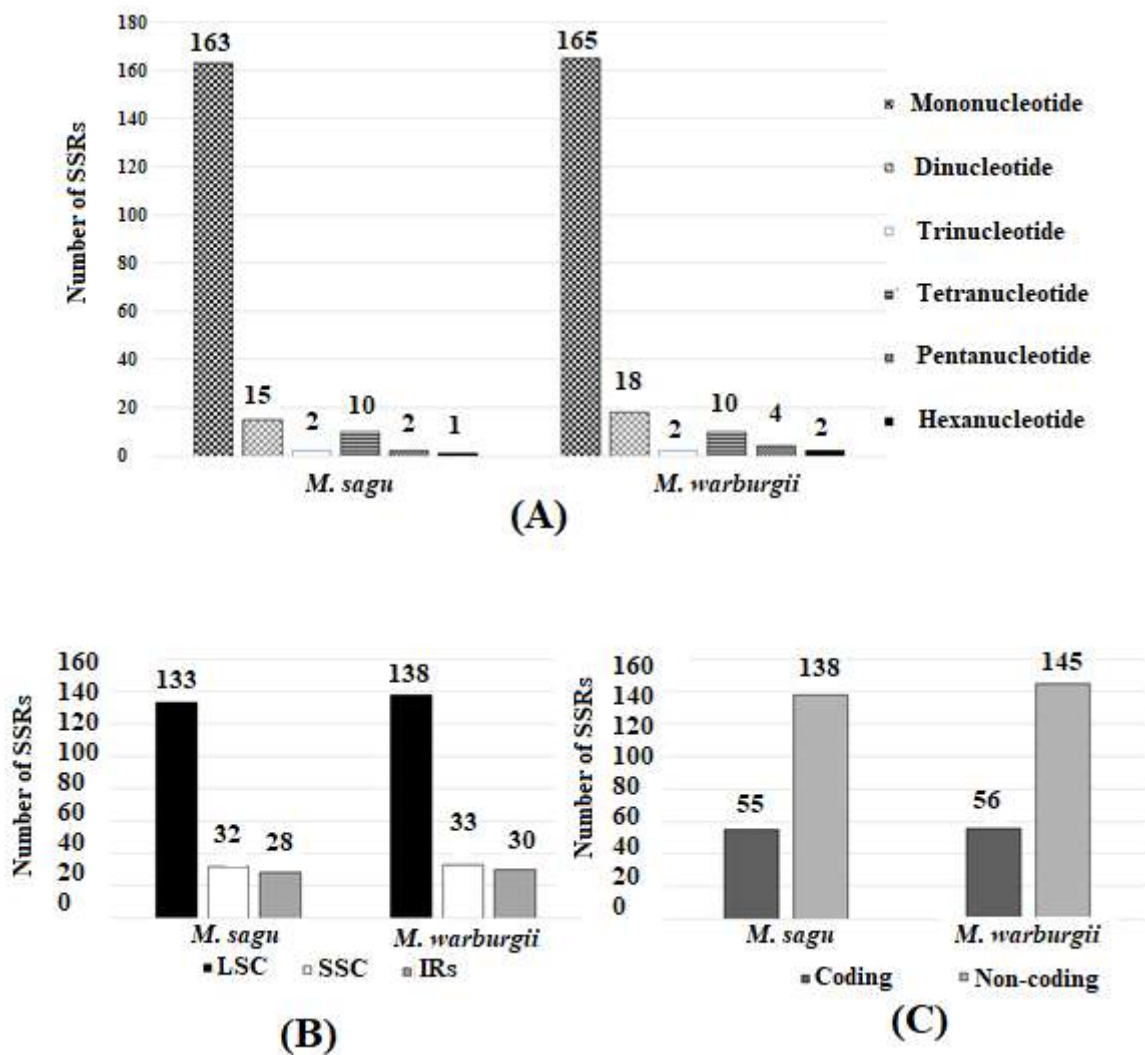


Figure 2.3: The distribution of SSRs in the chloroplast genomes of *Metroxylon sagu* and *Metroxylon warburgii*. (A) Number of varying SSR types in both *Metroxylon* species chloroplast genomes; (B) The distribution pattern of SSRs across different chloroplast genomic regions in both *Metroxylon* species genomes; (C) The distribution pattern of SSRs in coding and non-coding regions of the two *Metroxylon* species chloroplast genomes.

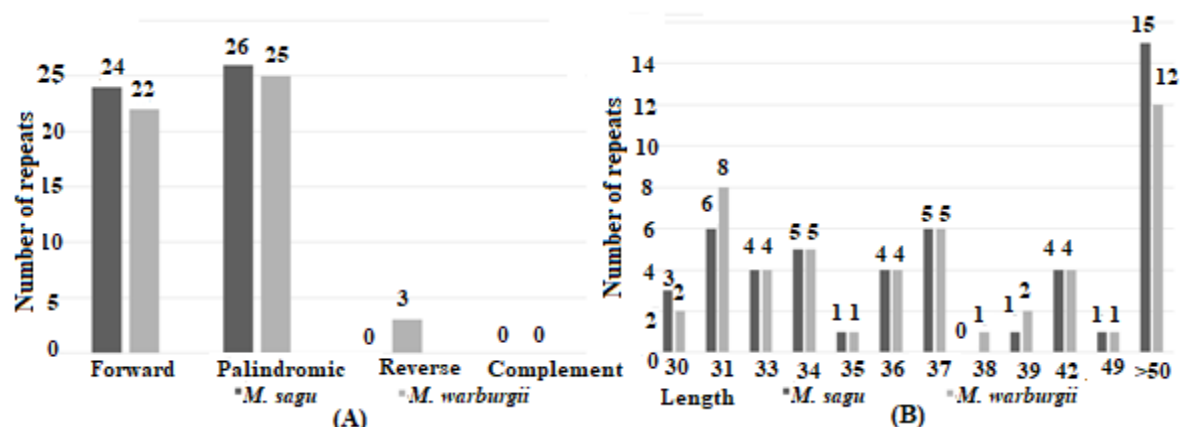


Figure 2.4: Long repeat sequences analysis on *M. sagu* and *M. warburgii* chloroplast genomes. (A) Long repeats types and frequency; (B) Long repeats frequency by length.

Table 2.5: The frequency of discovered microsatellite motifs across the *M. sagu* and *M. warburgii* chloroplast genomes.

Repeats	<i>M. sagu</i>	<i>M. warburgii</i>
A/T	156	158
AT/AT	13	16
C/G	7	7
AAAT/ATTT	5	5
AATG/ATTC	3	3
AG/CT	2	2
AAT/ATT	2	2
AGAT/ATCT	1	1
ATCC/ATGG	1	1
AAAGT/ACTTT	1	1
AACTT/AAGTT	1	1
AAAAGT/ACTTTT	1	1
AATAT/ATATT	0	1
AAAAAT/ATTTTT	0	1
AAAAGT/ACTTTT	0	1
Total Number of SSRs	193	201

2.3.4 Inverted Repeat Contraction and Expansion

The inverted repeat (IR) contraction and expansion were investigated at several locations such as both the single copy regions (SSC and LSC), two IRs (IRa and IRb) as well as the four junctions (SSC/IRa, IRa/LSC, LSC/IRb and IRb/SSC) among five palm relatives, namely *E. tristis*, *M. flexuosa*, *P. elata*, *M. warburgii* and *M. sagu* (Figure 2.5).

The IRs of all five palms are more conserved than other regions like the single copy regions. Only the *ycf1* gene in *E. tristis* is situated at the SSC/IRa junction whereas for others the *ycf1* gene is distanced 989 bp (for *M. flexuosa* and *P. elata*) and 971 bp (for both congeneric *Metroxylon* species) away from the junction. The IRa/LSC junction is free of any functional genes in all palms with *M. flexuosa* excluded (*rpl2* gene is right at the junction). Notably, the length and location of *rpl2-rps19* genes as well as *psbA* gene are highly similar with varying distances from the junction. In terms of distance, location and size of *rpl22-rps19* at the LSC/IRb junction, the *M. warburgii* exactly reflected that of *P. elata*. Moreover, the LSC/IRb junction was found to be highly conserved across all palm species except *M. flexuosa*. In *M. flexuosa*, the *rpl22* gene is displaced from the LSC/IRb junction, *rpl2* gene is positioned at the junction, and the LSC housed the *rps19* gene instead of IRb. All palms (except *E. tristis*) exhibited high degrees of conservation on *ycf1-ndhF* genes at the IRb/SSC junction with regard to gene lengths, location and distance. The *ndhF* gene is lost from the IRb/SSC junction in *E. tristis*, interestingly, a *rpl32* gene appeared in the SSC section and shared the same strand as the *ycf1* gene.

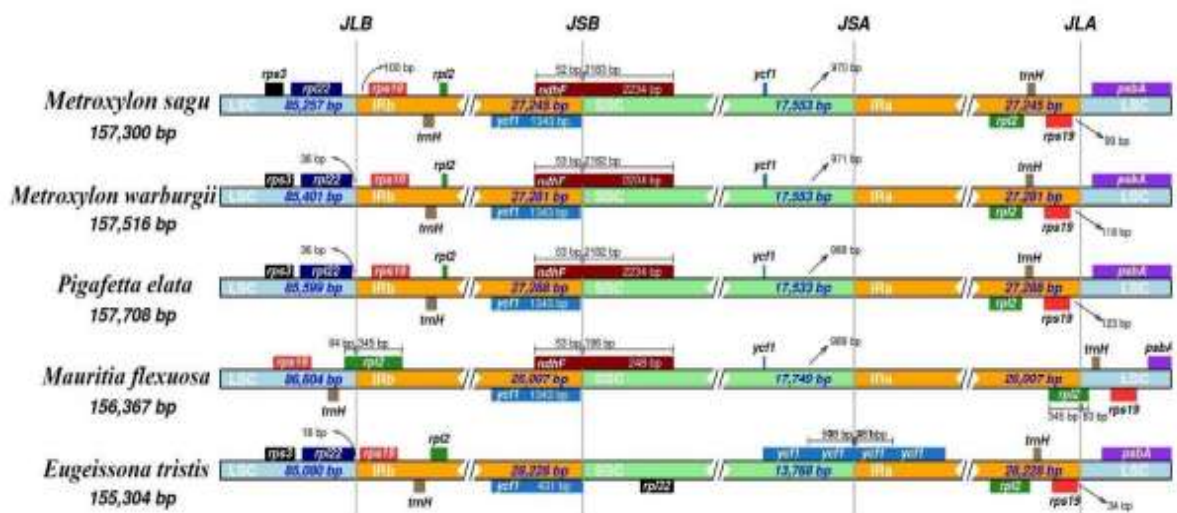


Figure 2.5: Comparison of IR/LSC and IR/SSC border positions across five Arecaceae family members chloroplast genomes. JLB: LSC/IRb junction; JSB: IRb/SSC junction; JSA: SSC/IRa junction; JLA: IRa/LSC junction.

2.3.5 Comparative Chloroplast Genome Analysis

The genome divergence was visualised across five different palm members via multiple sequence alignment using the sago palm chloroplast genome as the reference genome (Figure 2.6). Lower divergence was observed in the two IRs, in comparison to the single copy regions (LSC and SSC) across all five chloroplast genomes. The high divergence was indicated by the larger white spaces found across the alignment (Figure 2.6). The intergenic spacers in the SSC depicted lower divergence than that in the LSC. The low divergence (<50%) intergenic spacers identified were *ndhG-ndhI*, *rpl32-ccsA* as well as *psaC-ndhG*. On the other hand, the high divergence (>50%) intergenic spacers in the LSC were *accD-psaI*, *rps4-psbK*, *petA-psbJ*, *rps16-psbK*, *ndhC-atpE* and *trnH-psbA*. The conservation of non-coding regions across the five palms was low as compared to the coding regions. High divergence (>50%) occurred at coding regions like *ndhI*, *matK*, *ndhG*, *rps19*, *rpoA*, *ycf1* and *ndhA*.

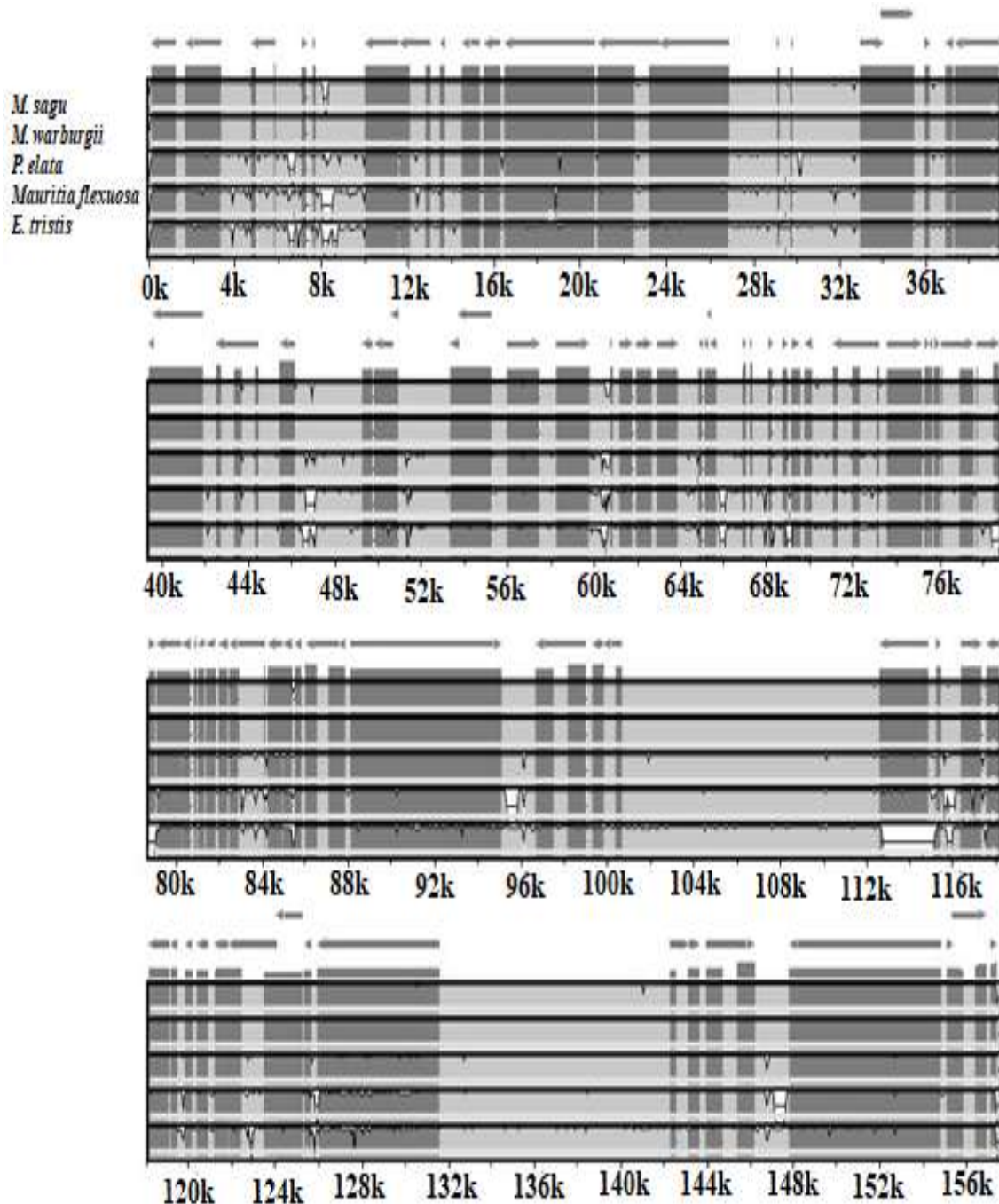


Figure 2.6: Comparison of chloroplast genomes across the five Arecaceae family members, with *M. sagu* as reference, employing mVISTA alignment program. Dark grey arrows represent gene orientations, dark grey bars indicate exons, light grey bars depict non-coding sequences and white spaces represent differences across the chloroplast genomes.

Furthermore, a survey of high nucleotide sequence variable regions using the DnaSP sliding window analysis was performed across the two *Metroxylon* congeneric

species (Figure 2.7A). The nucleotide variability was much higher in the SSC and LSC as to the IRs. The average nucleotide diversity value (Pi) determined was 0.00225. At the cut-off Pi value of 0.03, four mutational hotspots were revealed, namely *trnS-trnG*, *accD-ycf4*, *rpl22-rps19* and *trnH-rps19*. The former two mutational hotspots were from the LSC whereas the latter duos were from the IRs. In other words, the SSC is free from any presence of mutational hotpost at cut-off Pi value of 0.03.

The nucleotide variability investigation across five palm members as illustrated in Figure 2.7B, yielded an average Pi value of 0.01108 (a value greater than the comparison between two *Metroxylon* species). At cut-off Pi value of 0.04, a sum of nine mutational hotspots were unravelled across the five chloroplast genomes and none of them located in the IRs. The LSC region houses five of the identified mutational hotspots, namely *rps4-trnL*, *trnS-trnG*, *accD-trnL*, *rps16-psbK* and *rps4-trnL*. The remaining four were all from the SSC section: *adhA*, *ndhF*, *rps15-ycf1* and *ndhF-trnL*. These mutational hotspots are essential species level distinguishing tool that can be beneficial towards the study of palm taxonomy and evolution (Dong et al., 2012; Li et al., 2019).

The chloroplast genome of *M. warburgii* (157,516 bp) is 216 bp larger than that of *M. sagu* (157,300 bp). Apart from genome size disparity, their genome also differed in respect of a myriad of single nucleotide polymorphisms (SNPs). Approximately 63.4% of the total 164 SNPs discovered originated from the intergenic spacers. There are 60 SNPs found within the coding regions and majority (83.3%) of them are nonsynonymous (Table 2.6). The SNPs were found most abundantly within the *ycf1* gene across the two *Metroxylon* species, ahead of *accD* (eight SNPs) and *rpoC2* (five SNPs).

A total of 51 indels were unearthed from the comparison across the two *Metroxylon* congeneric species. The indels data displayed a balance between insertions and deletions

where 26 insertions and 25 deletions were recorded (Table 2.7). The largest indel marker (163 bp) is the intergenic spacer between *trnS* and *trnG* genes. Deletions were found in genes like *ycf3*, *rpoC1* and *clpP*, the *clpP* gene outshined the other two with a sum of three deletions observed. The *rpoC1* gene is the only gene encountering insertion.

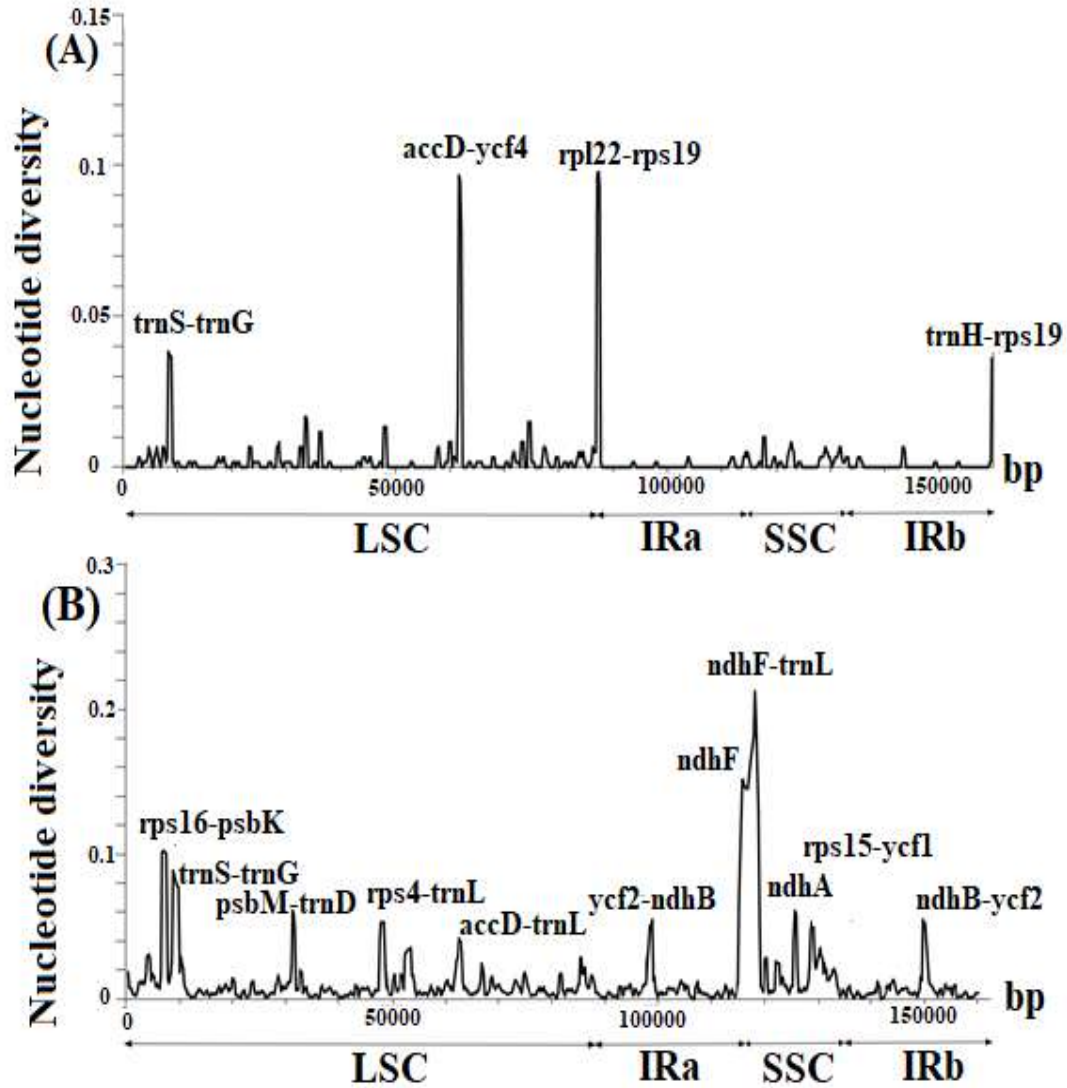


Figure 2.7: Sliding window analysis of the complete chloroplast genomes. Window length: 600 bp; Step size: 200 bp. X-axis: midpoint positions of a window; Y-axis: nucleotide diversity of each window. (A) Nucleotide diversity across *M. warburgii* and *M. sagu*. (B) Nucleotide diversity across *E. tristis*, both *Metroxylon* species, *P. elata* and *M. flexuosa*.

Table 2.6: SNP identified from the comparison between *M. warburgii* and *M. sagu* chloroplast genomes.

<i>M. sagu</i> nucleotide	<i>M.</i> <i>warburgii</i> nucleotide	Nucleotide differences	Gene/Intergenic	Mutation type	AA differences	Region
2651	2652	A-G	matK	nonsyn	F-L	LSC
2875	2876	A-C	matK	nonsyn	F-L	LSC
3705	3706	C-T	intergenic			LSC
4275	4276	G-T	intergenic			LSC
4655	4656	A-G	intergenic			LSC
4730	4731	C-A	intergenic			LSC
5020	5021	G-T	intergenic			LSC
5965	5965	G-A	intergenic			LSC
6089	6089	G-A	intergenic			LSC
6833	6833	T-C	intergenic			LSC
7703	7698	T-G	intergenic			LSC
7978	7973	G-T	intergenic			LSC
8086	8238	A-T	intergenic			LSC
8492	8646	A-G	intergenic			LSC
8902	9056	A-G	intergenic			LSC
9404	9558	C-A	intergenic			LSC
9813	9967	T-C	intergenic			LSC
12305	12459	C-T	intergenic			LSC
16452	16607	C-T	rpoC2	nonsyn	E-K	LSC
16880	17035	A-G	rpoC2	nonsyn	F-S	LSC
17482	17637	T-C	rpoC2	syn	L-L	LSC
17592	17747	T-G	rpoC2	nonsyn	K-Q	LSC
19518	19673	T-A	rpoC2	nonsyn	1-F	LSC
23219	23375	T-G	rpoC1	nonsyn	Y-S	LSC
23867	24023	G-C	rpoB	nonsyn	Q-E	LSC
27399	27555	A-T	intergenic			LSC
27400	27556	A-T	intergenic			LSC
27401	27557	A-C	intergenic			LSC
27474	27630	T-G	intergenic			LSC
27960	28116	T-G	intergenic			LSC
29019	29175	G-A	petN	nonsyn	V-I	LSC
29603	29754	G-A	psbM	syn	Y-Y	LSC
31573	31725	G-A	intergenic			LSC
31574	31726	T-C	intergenic			LSC
31581	31733	G-A	intergenic			LSC
31582	31734	T-C	intergenic			LSC
32511	32665	T-A	intergenic			LSC
32514	32675	C-A	intergenic			LSC
35269	35431	G-T	intergenic			LSC

Table 2.6 continued

35270	35432	G-C	intergenic			LSC
35271	35433	T-C	intergenic			LSC
35272	35434	G-C	intergenic			LSC
35273	35435	G-A	intergenic			LSC
35274	35436	G-C	intergenic			LSC
35275	35437	A-C	intergenic			LSC
36758	36926	T-G	intergenic			LSC
41915	42083	C-A	intergenic			LSC
42821	42989	C-T	intergenic			LSC
43069	43236	G-T	intergenic			LSC
43315	43482	A-G	ycf3	syn	S-S	LSC
44085	44261	G-T	ycf3			LSC
44139	44315	T-G	ycf3			LSC
46088	46258	C-A	intergenic			LSC
46731	46901	T-G	intergenic			LSC
56456	56650	T-G	rbcL	nonsyn	L-V	LSC
56473	56667	G-A	rbcL	syn	A-A	LSC
56474	56668	C-A	rbcL	nonsyn	L-I	LSC
56594	56728	A-G	rbcL	nonsyn	I-V	LSC
57857	58053	C-A	intergenic			LSC
57858	58054	C-A	intergenic			LSC
58083	58278	G-T	accD	nonsyn	L-F	LSC
58674	58869	A-T	accD	nonsyn	K-N	LSC
58675	58870	T-A	accD	nonsyn	S-R	LSC
58676	58871	C-T	accD	nonsyn	S-R	LSC
58677	58872	C-A	accD	nonsyn	S-R	LSC
58708	58903	A-C	accD	nonsyn	K-Q	LSC
59468	59663	C-T	accD	nonsyn	A-V	LSC
59527	59722	G-T	accD	nonsyn	L-Q	LSC
60119	60313	C-G	intergenic			LSC
62202	62425	G-T	cemA	nonsyn	L-F	LSC
63656	63879	C-G	intergenic			LSC
64144	64367	T-G	intergenic			LSC
66378	66601	T-G	intergenic			LSC
66435	66658	C-A	intergenic			LSC
70865	71079	C-A	intergenic			LSC
75614	75834	T-A	psbH	nonsyn	F-L	LSC
75734	75954	T-C	psbH	syn	N-N	LSC
75814	76034	T-G	intergenic			LSC
76091	76311	A-G	intergenic			LSC
77867	78087	A-T	intergenic			LSC
77868	78088	G-T	intergenic			LSC
80321	80541	T-C	rps11	syn	R-R	LSC

Table 2.6 continued

81592	81812	C-T	intergenic			LSC
81675	81895	G-T	intergenic			LSC
82056	82276	G-A	rpl14	nonsyn	A-V	LSC
82319	82539	G-T	rpl16	nonsyn	S-E	LSC
82320	82540	A-C	rpl16	nonsyn	S-E	LSC
82892	83112	G-A	intergenic			LSC
83975	84195	A-G	intergenic			LSC
84380	84600	C-T	rps3	nonsyn	R-K	LSC
84481	84701	T-C	rps3	nonsyn	G-R	LSC
85108	85328	C-T	rpl22	nonsyn	R-K	LSC
85204	85424	T-G	intergenic			LSC
85209	85429	G-T	intergenic			LSC
85210	85430	A-T	intergenic			LSC
85211	85431	A-T	intergenic			LSC
85212	85432	A-C	intergenic			LSC
85217	85437	A-C	intergenic			LSC
85220	85440	A-T	intergenic			LSC
85223	85443	A-T	intergenic			LSC
85229	85449	G-T	intergenic			LSC
85230	85450	A-T	intergenic			LSC
85231	85451	A-C	intergenic			LSC
85306	85483	G-C	intergenic			LSC
85313	85490	A-C	intergenic			LSC
85576	85737	C-A	rps19	nonsyn	I-M	IR
91749	91910	G-T	ycf2	nonsyn	L-F	IR
95893	96054	T-G	intergenic			IR
109929	110108	C-A	intergenic			IR
112151	112330	G-C	ycf1	nonsyn	R-T	IR
112152	112331	G-C	ycf1	nonsyn	R-T	IR
112715	112894	C-T	ndhF	nonsyn	G-D	SSC
112937	113116	G-T	ndhF	nonsyn	P-Q	SSC
114858	115037	G-A	intergenic			SSC
115795	115973	G-A	intergenic			SSC
117233	117411	A-C	intergenic			SSC
117234	117412	T-A	intergenic			SSC
117235	117413	T-A	intergenic			SSC
117236	117414	G-T	intergenic			SSC
117331	117509	C-A	intergenic			SSC
117485	117663	C-G	ndhD	nonsyn	V-L	SSC
118509	118687	G-A	ndhD	syn	Y-Y	SSC
120137	120315	T-A	intergenic			SSC
120331	120509	A-G	ndhG	nonsyn	I-T	SSC
120425	120603	G-A	ndhG	nonsyn	P-S	SSC

Table 2.6 continued

120754	120932	A-T	intergenic			SSC
120760	120938	T-A	intergenic			SSC
120766	120944	A-T	intergenic			SSC
121989	122167	G-T	ndhA	syn	I-I	SSC
125500	125679	T-A	intergenic			SSC
126400	126579	T-C	ycf1	nonsyn	R-G	SSC
126755	126934	T-A	ycf1	nonsyn	K-N	SSC
127392	127571	A-C	ycf1	nonsyn	L-W	SSC
128071	128250	A-G	ycf1	nonsyn	Y-H	SSC
128528	128707	C-A	ycf1	nonsyn	K-N	SSC
128655	128834	G-T	ycf1	nonsyn	A-E	SSC
129063	129242	C-T	ycf1	nonsyn	G-E	SSC
129119	129298	G-T	ycf1	syn	L-L	SSC
129310	129489	T-C	ycf1	nonsyn	M-V	SSC
129374	129553	G-A	ycf1	syn	N-N	SSC
129865	130044	C-A	ycf1	nonsyn	D-Y	SSC
130406	130585	C-G	ycf1	nonsyn	R-T	SSC
130407	130586	C-G	ycf1	nonsyn	R-T	SSC
132540	132719	G-C	intergenic			IR
132629	132808	G-T	intergenic			IR
133024	133203	G-A	intergenic			IR
146665	146862	A-C	intergenic			IR
150809	151006	C-A	ycf2	nonsyn	L-F	IR
156982	157179	G-T	rps19	nonsyn	M-I	IR
157229	157426	T-G	intergenic			IR
157234	157442	C-A	intergenic			IR
157243	157458	T-A	intergenic			IR
157245	157460	T-A	intergenic			IR
157250	157465	T-G	intergenic			IR
157251	157466	T-A	intergenic			IR
157252	157467	C-A	intergenic			IR
157258	157473	T-A	intergenic			IR
157261	157476	T-G	intergenic			IR
157264	157479	T-G	intergenic			IR
157269	157484	T-A	intergenic			IR
157270	157485	T-A	intergenic			IR
157271	157486	T-A	intergenic			IR
157272	157487	C-A	intergenic			IR
157277	157492	A-C	intergenic			IR

Table 2.7: Indel comparison across both *Metroxylon* species chloroplast genomes. Ms: *M. sagu*; Mw: *M. warburgii*.

Indel type	Ms indel start	Ms indel end	Mw indel start	Mw indel end	Gene/Intergenic	Region	Indel length
Deletion	78	78	78	79	intergenic	LSC	1
Deletion	6650	6650	6650	6651	intergenic	LSC	1
Deletion	8073	8073	8068	8069	intergenic	LSC	1
Deletion	8082	8082	8072	8235	intergenic	LSC	163
Deletion	8093	8093	8246	8248	intergenic	LSC	2
Deletion	14021	14021	14175	14177	intergenic	LSC	2
Deletion	22476	22476	22632	22633	rpoC1	LSC	1
Deletion	22792	22792	22948	22949	rpoC1	LSC	1
Deletion	32500	32500	32653	32655	intergenic	LSC	2
Deletion	32512	32512	32667	32674	intergenic	LSC	7
Deletion	35251	35251	35413	35414	intergenic	LSC	1
Deletion	36160	36160	36323	36329	intergenic	LSC	6
Deletion	43538	43538	43706	43715	ycf3	LSC	9
Deletion	46746	46746	46922	46942	intergenic	LSC	20
Deletion	57862	57862	58058	58060	intergenic	LSC	2
Deletion	60367	60367	60562	60612	intergenic	LSC	50
Deletion	71463	71463	71678	71679	clpP	LSC	1
Deletion	71722	71722	71938	71943	clpP	LSC	5
Deletion	72277	72277	72499	72504	clpP	LSC	5
Deletion	101709	101709	101871	101889	intergenic	IR	18
Deletion	115623	115623	115802	115803	intergenic	SSC	1
Deletion	125448	125448	125627	125628	intergenic	SSC	1
Deletion	140830	140830	141010	141028	intergenic	IR	18
Deletion	157233	157233	157431	157442	intergenic	IR	11
Deletion	157242	157242	157451	157458	intergenic	IR	7
Insertion	4699	4700	4700	4700	intergenic	LSC	1
Insertion	6543	6544	6543	6543	intergenic	LSC	1
Insertion	7045	7050	7045	7045	intergenic	LSC	5
Insertion	8044	8049	8038	8038	intergenic	LSC	5
Insertion	13671	13672	13824	13824	intergenic	LSC	1
Insertion	22491	22492	22646	22646	rpoC1	LSC	1
Insertion	29490	29494	29645	29645	intergenic	LSC	4
Insertion	42841	42842	43008	43008	intergenic	LSC	1
Insertion	46013	46019	46188	46188	intergenic	LSC	6
Insertion	48626	48628	48821	48821	intergenic	LSC	2
Insertion	58041	58042	58237	58237	intergenic	LSC	1
Insertion	59622	59623	59816	59816	intergenic	LSC	1
Insertion	60462	60483	60705	60705	intergenic	LSC	21

Table 2.7 continued

Insertion	68108	68110	68330	68330	intergenic	LSC	2
Insertion	70098	70105	70318	70318	intergenic	LSC	7
Insertion	72991	72996	73216	73216	intergenic	LSC	5
Insertion	74995	74996	75215	75215	intergenic	LSC	1
Insertion	85236	85239	85455	85455	intergenic	LSC	3
Insertion	85242	85250	85458	85458	intergenic	LSC	8
Insertion	85254	85256	85462	85462	intergenic	LSC	2
Insertion	85259	85268	85465	85465	intergenic	LSC	9
Insertion	85276	85291	85473	85473	intergenic	LSC	15
Insertion	85298	85304	85480	85480	intergenic	LSC	6
Insertion	85309	85325	85485	85485	intergenic	LSC	16
Insertion	114948	114949	115126	115126	intergenic	SSC	1
Insertion	115638	115639	115816	115816	intergenic	SSC	1

2.3.6 RNA Editing

RNA editing events involve the alteration of nucleotide cytidine to uridine in the transcribed chloroplastic and mitochondrial RNAs and they are capable of orchestrating the nucleotide number and function (Larkin, 2019). With reference to the *C. nucifera* chloroplast genome, a sum of 126 RNA editing sites had been discovered within the sago palm chloroplast genome (Table 2.8). Amino acid changes were observed across all RNA editing events identified in the sago palm chloroplast genome and *petD* gene is the only gene encountering a stop codon subsequent to an amino acid change. Most of the RNA editing events (76.4%) were predicted to take place at the second position. The RNA editing events of sago palm that happen at the first position was enumerated as 32 sites (25.4%). Interestingly, none was expected to occur at the third position.

Serine to leucine was the most common changes of amino acid, with 41 sites recorded (32.5%). This was followed proline to leucine (11.9%) and histidine to tyrosine (11.1%). At the gene level, RNA editing events were found in a total of 37 genes in the sago palm chloroplast genome (Figure 2.8). There are 12 genes (32.4%) with one RNA

editing sites and 10 genes (27%) with two RNA editing sites. The gene with the most abundant RNA editing sites is the *ycf1* gene where 19 sites were detected. The first and second runners-up are *ndhB* and *matK* genes as 12 and 10 sites were unravelled, respectively.

Table 2.8: RNA editing sites in sago palm chloroplast transcripts.

Gene	Base	Amino acid length	Triplet position	Bases	Codon	Amino acid change
matK	271	91	1	C-U	CAU-UAU	H-Y
	734	245	2	C-U	UCU-UUU	S-F
	824	275	2	C-U	UCC-UUC	S-F
	943	315	1	C-U	CAU-UAU	H-Y
	1097	366	2	C-U	ACU-AUU	T-I
	1267	423	1	C-U	CAC-UAC	H-Y
	1321	441	1	C-U	CUU-UUU	L-F
	1361	454	2	C-U	ACG-AUG	T-M
	1364	455	2	C-U	GCA-GUA	A-V
	1534	512	1	C-U	CAU-UAU	H-Y
rps16	155	52	2	C-U	GCU-GUU	A-V
atpA	914	305	2	C-U	UCA-UUA	S-L
	1148	383	2	C-U	UCA-UUA	S-L
atpF	92	31	2	C-U	CCA-CUA	P-L
atpI	629	210	2	C-U	UCA-UUA	S-L
rps2	134	45	2	C-U	ACA-AUA	T-I
	242	81	2	C-U	GCG-GUG	A-V
	248	83	2	C-U	UCA-UUA	S-L
	400	134	1	C-U	CCC-UCC	P-S
rpoC2	1201	401	1	C-U	CCA-UCA	P-S
	2309	770	2	C-U	UCG-UUG	S-L
	2662	888	1	C-U	CCC-UCC	P-S
	3701	1234	2	C-U	UCA-UUA	S-L
rpoC1	41	14	2	C-U	CCA-CUA	P-L
	511	171	1	C-U	CGG-UGG	R-W
	617	206	2	C-U	UCA-UUA	S-L
	1304	435	2	C-U	CCG-CUG	P-L
rpoB	467	156	2	C-U	UCA-UUA	S-L
	545	182	2	C-U	UCA-UUA	S-L
	560	187	2	C-U	UCG-UUG	S-L

Table 2.8 continued

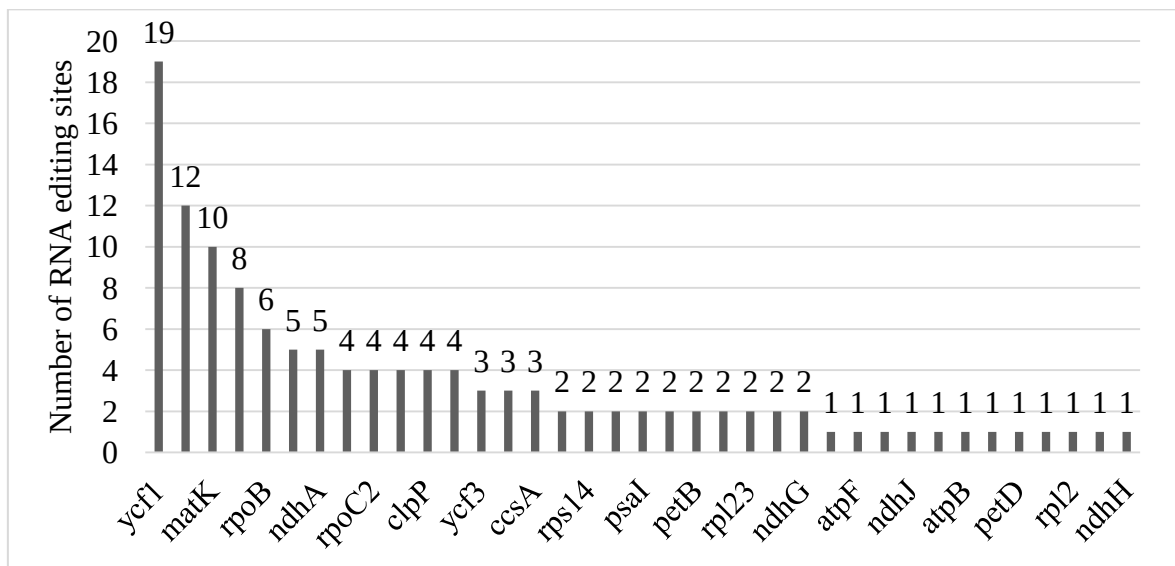
	627	206	2	C-U	CCG-CUG	P-L
	1994	665	2	C-U	UCU-UUU	S-F
	2420	807	2	C-U	UCA-UUA	S-L
rps14	80	27	2	C-U	UCA-UUA	S-L
	149	50	2	C-U	CCA-CUA	P-L
ycf3	44	15	2	C-U	UCU-UUU	S-F
	185	62	2	C-U	ACG-AUG	T-M
	191	64	2	C-U	CCG-CUG	P-L
ndhJ	131	44	2	C-U	UCA-UUA	S-L
ndhK	131	44	2	C-U	UCG-UUG	S-L
ndhC	311	104	2	C-U	CCA-CUA	P-L
	323	108	2	C-U	UCA-UUA	S-L
atpB	1184	395	2	C-U	UCA-UUA	S-L
accD	794	265	2	C-U	UCG-UUG	S-L
	1157	386	2	C-U	UCA-UUA	S-L
	1385	462	2	C-U	CCA-CUA	P-L
	1403	468	2	C-U	CCU-CUU	P-L
	1415	472	2	C-U	GCU-GUU	A-V
psaI	80	27	2	C-U	UCU-UUU	S-F
	85	29	1	C-U	CAU-UAU	H-Y
ycf4	41	14	2	C-U	ACG-AUG	T-M
	254	85	2	C-U	UCA-UUA	S-L
cemA	34	12	1	C-U	CCC-UCC	P-S
clpP	82	28	1	C-U	CAU-UAU	H-Y
	557	186	2	C-U	GCU-GUU	A-V
	559	187	1	C-U	CAU-UAU	H-Y
	608	203	2	C-U	UCC-UUC	S-F
petB	418	140	1	C-U	CGG-UGG	R-W
	611	204	2	C-U	CCA-CUA	P-L
petD	481	161	1	C-U	CAA-UAA	Q-*
rpoA	200	67	2	C-U	UCU-UUU	S-F
	368	123	2	C-U	UCA-UUA	S-L
	830	277	2	C-U	UCA-UUA	S-L
rps8	182	61	2	C-U	UCA-UUA	S-L
rps3	470	157	2	C-U	ACA-AUA	T-I
	583	195	1	C-U	CAU-UAU	H-Y
rpl2	2	1	2	C-U	ACG-AUG	T-M
rpl23	71	24	2	C-U	UCU-UUU	S-F
	89	30	2	C-U	UCA-UUA	S-L
ycf2	5561	1854	2	C-U	GCA-GUA	A-V
	6026	2009	2	C-U	UCA-UUA	S-L
ndhB	149	50	2	C-U	UCA-UUA	S-L
	467	156	2	C-U	CCA-CUA	P-L

Table 2.8 continued

ndhF	542	181	2	C-U	ACG-AUG	T-M
	586	196	1	C-U	CAU-UAU	H-Y
	704	235	2	C-U	UCC-UUC	S-F
	737	246	2	C-U	CCA-CUA	P-L
	830	277	2	C-U	UCA-UUA	S-L
	836	279	2	C-U	UCA-UUA	S-L
	1112	371	2	C-U	UCA-UUA	S-L
	1193	398	2	C-U	UCA-UUA	S-L
	1255	419	1	C-U	CAU-UAU	H-Y
	1481	494	2	C-U	CCA-CUA	P-L
	62	21	2	C-U	UCA-UUA	S-L
	290	97	2	C-U	UCA-UUA	S-L
	442	148	1	C-U	CAU-UAU	H-Y
	1787	596	2	C-U	ACA-AUA	T-I
rpl32	19	7	1	C-U	CGU-UGU	R-C
ccsA	11	4	2	C-U	GCG-GUG	A-V
ndhD	94	32	1	C-U	CAC-UAC	H-Y
	299	100	2	C-U	ACA-AUA	T-I
	2	1	2	C-U	ACG-AUG	T-M
	59	20	2	C-U	UCA-UUA	S-L
	383	128	2	C-U	CCA-CUA	P-L
ndhG	674	225	2	C-U	UCG-UUG	S-L
	947	316	2	C-U	ACA-AUA	T-I
	1193	398	2	C-U	UCA-UUA	S-L
	1310	437	2	C-U	UCA-UUA	S-L
	1319	440	2	C-U	ACG-AUG	T-M
	259	87	1	C-U	CCC-UCC	P-S
	347	116	2	C-U	CCA-CUA	P-L
	50	17	2	C-U	UCG-UUG	S-L
	476	159	2	C-U	UCA-UUA	S-L
	566	189	2	C-U	UCA-UUA	S-L
ndhA	863	288	2	C-U	ACU-AUU	T-I
	961	321	1	C-U	CCU-UCU	P-S
	545	182	2	C-U	UCU-UUU	S-F
	737	246	2	C-U	ACG-AUG	T-M
	838	280	1	C-U	CCC-UCC	P-S
	1543	515	1	C-U	CGU-UGU	R-C
	1642	548	1	C-U	CUU-UUU	L-F
	1712	571	2	C-U	UCA-UUA	S-L
	1835	612	2	C-U	ACU-AUU	T-I
	1840	614	1	C-U	CCA-UCA	P-S
ycf1	2450	817	2	C-U	UCG-UUG	S-L
	2482	828	1	C-U	CAC-UAC	H-Y

Table 2.8 continued

2981	994	2	C-U	UCC-UUC	S-F
3395	1132	2	C-U	ACA-AUA	T-I
3664	1222	1	C-U	CCU-UCU	P-S
3742	1248	1	C-U	CAU-UAU	H-Y
3866	1289	2	C-U	ACG-AUG	T-M
4598	1533	2	C-U	UCA-UUA	S-L
4607	1536	2	C-U	UCC-UUC	S-F
4958	1653	2	C-U	ACA-AUA	T-I
5134	1712	1	C-U	CCC-UCC	P-S
5191	1731	1	C-U	CUU-UUU	L-F

**Figure 2.8:** The distribution of RNA editing sites across sago palm chloroplast genes.

2.3.7 Phylogenetic Analysis

The maximum likelihood phylogenetic tree was constructed based on all protein-coding genes of 39 palms and 2 outgroups. Figure 2.9 depicted five distinguishing major clades which represents subfamilies like Arecoideae, Nypoideae, Calamoideae, Coryphoideae and Ceroxyloideae, with miniature exceptions. The one outstanding clade encompassing *Podococcus barteri* and *Tahina spectabilis* is exceptionally strong and both

of them are situated distant from their expected subfamily classification, Arecoideae and Coryphoideae respectively.

The sago palm shared the major Calamoideae subfamily clade with other palm members such as *M. flexuosa*, *M. warburgii*, *P. elata*, *Calamus caryotoides*, *Eremospatha macrocarpa*, *E. tristis* and *Salacca ramosiana* (Yang et al., 2010; Uthaipaisanwong et al., 2012; Huang et al., 2013; Barrett et al., 2016; Lim et al., 2020). Zooming in to the Calameae tribe where the sago palm resides, the tribe is strong with 100% bootstrap values (Figure 2.9) across all branches, consisting of *M. warburgii*, *S. ramosiana*, *C. caryotoides*, *P. elata* and *M. sagu*. At the neighbouring tribes within the same subfamily clade, *M. flexuosa* and *E. macrocarpa* form the Lepidocaryeae tribe whereas *E. tristis* is the only member of Eugeissoneae tribe.

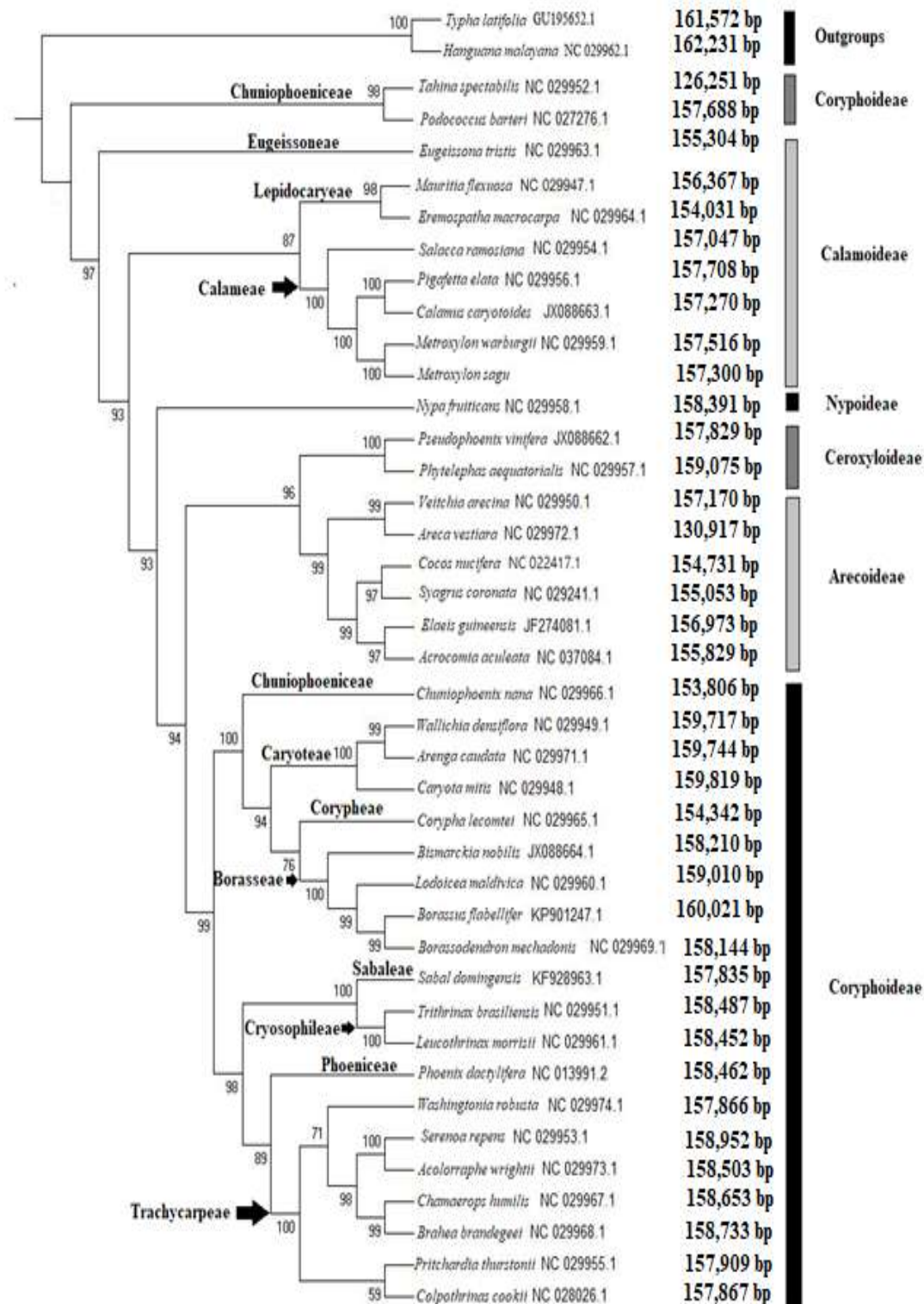


Figure 2.9: The maximum likelihood phylogenetic tree of Arecaceae family members inferred from chloroplast protein-coding genes, with two outgroup gymnosperms (*Typha latifolia* and *Hanguana malayana*) included into analysis. The labels on the right represent the subfamilies of the Arecaceae family whereas the labels on the left represent Calamoid and Coryphoid tribes, with their respective genome sizes respectively.

2.4 Discussion

2.4.1 Chloroplast Genome Organisation

In this study, the complete chloroplast genome sequence of sago palm was characterised as a closed circular quadripartite conformation consisting of LSC, SSC and two IRs, a conserved structure as seen in other members of palm family like *E. guineensis*, *M. warburgii*, *Colpothrinax cookii*, *M. flexuosa*, *P. elata*, *E. trisits*, *S. ramosiana*, *C. caryotoides* and *P. dactylifera* (Yang et al., 2010; Uthaipaisanwong et al., 2012; Huang et al., 2013; Barrett et al., 2016). There is a total of 113 unique genes and 23 duplicative genes residing in the sago palm chloroplast genome. The sum of 29 tRNA genes, four ribosomal RNA genes as well as 80 functional genes in the sago palm chloroplast genome is highly conserved across majority of the Arecaceae family members. Some genes (*ycf1*, *accD* and *ycf2*) are found exclusively in palm members like *M. sagu*, *E. guineensis* and *M. warburgii* (Uthaipaisanwong et al., 2012; Barrett et al., 2016) but not found in chloroplast genomes of Poaceae (Guisinger et al., 2010). Similarly, the *infA* gene that is absent in duckweed and dicots (Mardanov et al., 2008), was found to appear in most palm family members, including the sago palm.

2.4.2 Nucleotide Variation Analysis

The nucleotide variation analysis on various genomic features such as SNPs, indels, microsatellites and long repeats can greatly facilitate the improvement of resolution to plant phylogeny at low taxonomic levels (Dong et al., 2012). In the present study, five out of the eleven sites with high divergence identified across the selected five palm chloroplasts, namely *accD-trnL*, *ndhF*, *trnS-trnG* and *ndhF-trnL*, are well-known powerful

markers in DNA barcoding and plant molecular systematics (Bakker et al., 2000; Taberlet et al., 2007; Dong et al., 2012; Silva et al., 2016; Li et al., 2018). Furthermore, the four mutational hotspots found across the two *Metroxylon* congeneric species (*trnH-rps19*, *accD-ycf4*, *trnS-trnG* and *rpl22-rps19*) are potential markers vital for species identification across the *Metroxylon* genus. Particularly, two of the abovementioned markers, *accD-ycf4* and *trnS-trnG*, are well-established species markers for a myriad of recently published plant chloroplast researches, to name a few, *Hibiscus* (Abdullah et al., 2019), *Fritillaria* (Li et al., 2018), *Populus* (Zong et al., 2019), *Boswellia* (Al-Harrasi et al., 2019), *Angelica* (Park et al., 2019) and *Aconitum* (Park et al., 2017).

Moreover, the microsatellites and long repeats data of the sago palm provide additional windows to tap into the genetic profile of this pivotal palm species. In the present study, the number of microsatellites as well as the composition are highly conserved across the two *Metroxylon* species tested. A total of 201 and 193 microsatellites were discovered from the chloroplast genomes of *M. warburgii* and *M. sagu* respectively. Looking at the composition, 14.5% of microsatellites are from IRs, 16.6% are from SSC and 68.9% are from LSC for sago palm while 14.9% of SSRs are from IRs, 16.4% are from SSC and 68.7% are from LSC for *M. warburgii*. In general, the microsatellites are more abundant in SSC than in IRs for both *Metroxylon* species but that is not the case for other plants like silver spot, oil palm, galangal and fameflower where the condition is vice versa (Uthaipaisanwong et al., 2012; Liu et al., 2018; Li et al., 2019). The mononucleotide repeats dominated the microsatellite population for both *Metroxylon* species, a similar phenomenon observed across other plants like oil palm, fameflower, silver spot and galangal but in wild rice (*Oryza minuta*) the dinucleotide number outshined that of the other SSRs (Uthaipaisanwong et al., 2012; Asaf et al., 2017; Liu et al., 2018; Li et al.,

2019). The forward long repeats were surpassed by palindromic long repeats and complement long repeat is nowhere to be found across the two *Metroxylon* congeneric species. On the contrary, the exact converse situation was detected in galangal and silver spot (Liu et al., 2018; Li et al., 2019). These data are pivotal steppingstones towards the establishment of genetic and taxonomical studies of various palms.

The nucleotide substitution events investigated in this study revealed a sum of 164 SNPs across *M. sagu* and *M. warburgii*. This number is fairly conserved among other plant pairs like *Machilus yunnanensis* and *M. balansae* (231 SNPs) as well as *O. nivara* and *O. sativa* (159 SNPs), but seems rather miniature in comparison to that of *Citrus sinensis* and *C. aurantiifolia* (330 SNPs), *Fagopyrum dibotrys* and *F. tataricum* (317 SNPs) as well as *Kaempferia galanga* and *K. elegans* (536 SNPs) (Shahid Masood et al., 2004; Su et al., 2014; Song et al., 2015; Wang et al., 2017; Li et al., 2019). Furthermore, the 51 indels scored between *M. sagu* and *M. warburgii* bear a close resemblance to that between *Aconitum pseudolaeve* and *A. longecassidatum* (53 indels) as well as between *M. yunnanensis* and *M. balansae* (65 indels), however deemed less as to between *K. elegans* and *K. galanga* (107 indels) as well as *Panax notoginseng* and *P. ginseng* (156 indels) (Dong et al., 2014; Song et al., 2015; Park et al., 2017; Li et al., 2019). The *clpP*, *ycf3* and *rpoC1* genes encountered deletions within their sequences and this phenomenon is common as they are known to be highly varied among other plant species (Dong et al., 2012; Wang et al., 2017). The largest indel between *M. sagu* and *M. warburgii* is 163 bp in length and situated between *trnS* and *trnG* genes, an analogous condition was observed between *Solanum bulbocastanum* (wild potato) and *S. tuberosum* (cultivated potato) but the length is slightly greater (241 bp) (Chung et al., 2006). Besides, the *trnS-trnG* intergenic deletion region transcended all other heterogenous regions as the highest

diverged (12%) section among new world Lecythidaceae species (Hamilton et al., 2003). Mayonde et al. (2015) had also employed this region to successfully unravel the genetic taxonomy of *Tamarix* species and their putative hybrids.

2.4.3 Phylogenetic Analysis

The maximum likelihood phylogenetic tree of the 39 palms depicted a high accuracy in terms of the correlation between subfamily classification and branch assignment with some exceptions (*P. barteri* and *T. spectabilis*). Barrett et al. (2016) had performed phylogenetic analysis across all palms available at the time of publication utilizing the complete chloroplast genome sequences. The resemblance of the tree constructed in this study to that built by Barrett et al. (2016) is high with the common abovementioned exceptions and with the sago palm being included into analysis.

According to Barrett et al. (2016), *P. barteri* contains partial inverted repeats and was not included into the phylogenetic tree by Barrett et al. (2016). This species has not been characterised and sequenced to date. As for *T. spectabilis*, Barrett et al. (2016) has identified that this species lacks an inverted repeat region but instead conserved a 159 bp *ndhB* duplicative fragment. This has been taken into account by Barrett et al. (2016) when performing the phylogenetic analysis. Congruent topology can be observed across subfamilies Arecoideae, Nypoideae, Ceroxyloideae, Coryphoideae, Calamoideae as well as Calamoid tribes (Trachycarpeae, Sabaleae, Borasseae, Caryoteae, Phoeniceae, Corypheae, Cryosophileae and Chuniophoeniceae). The complete chloroplast genome of the sago palm would aid greatly in terms of future research on its taxonomical status and chloroplast metabolic pathway comprehension. All in all, these outcomes serve as vital foundation for future molecular works.

CHAPTER 3

ORGANELLAR GENOME PROFILES ACROSS ORGANS, GROWTH STAGES, PHENOTYPES AND LOCALITIES

3.1 Introduction

The organellar genomes consisting of the mitogenome and plastogenome (or plastome) in plants are constantly subjected to attacks by reactive metabolic by-products and UV lights during their involvement in biological processes such as photosynthesis and respiration (Boesch et al., 2010; Alexeyev et al., 2013; Kumar et al., 2014). Due to the lack of efficient DNA repair mechanism as seen in their nuclear counterpart, the highly polyploid nature of the organellar genome favour the degradation of damaged molecules in order to avoid cancer progression due to mutation (Liu & Demple, 2010). This explains for the organellar genome copy number variations across different stages of development of plants and when subjected to environmental stress and stimuli (Kumar et al., 2014).

The mitogenome stoichiometric variations can impact the phenotypes of plant greatly and severely, even causing disease phenotypes such as cytoplasmic male sterility (Li et al., 1988; Arrieta-Montiel et al., 2001), leaf shape variations (Li et al., 1988) as well as plant growth speed (Ma & Li, 2015). Under the conditions that does not lead to diseases, the high copy number of organellar DNA is a strong indication of great demand for organellar ribosomes and the possible way to fulfil this demand is no other than the elevation of rRNA level from the event of genome multiplication (Bendich, 1987). The copy number variation in plants is one of the major drivers in phenotypic variations, orchestrating heritable dissimilar traits between individuals (Zhang et al., 2009; Iskow et al., 2012). In maize, the organellar DNA copy number can fluctuate up to 20-fold under the

influence of different light exposures at different tissues as the copy number becomes the limiting factor for the expression of some chloroplast genes (Oldenburg et al., 2006; Zheng et al., 2011; Udy et al., 2012).

Preceding organellar genome investigation research has been concentrating on prevalent plant species like tobacco (Ma & Li, 2015), maize (Kumar et al., 2014) and *Arabidopsis* (Li et al., 2006). All of them have come to the same conclusion that organellar profiles do fluctuate across different developmental stages and there is a disparity between smaller and higher plants in terms of their mechanism of maintain the organellar genome copy numbers (Li et al., 2006). This leads to the need for the investigation of the copy number and integrity of organellar genome of sago palm in order to further understand their correlation with growth and development of the sago palm. The identification of the organellar DNA copy number at different growth stages, phenotypes and localities of the sago palm is essential as it allows for the detailed overview of the starch metabolism in this unique plant (Zhang et al., 2009; Iskow et al., 2012; Gualberto & Newton, 2017) where starch is stored in its stem (Dransfield, 1976; Ruddle, 1976; Tan, 1983; Chew et al., 1999). It is still unknown whether the organellar DNA copy number will change at different developmental stages of sago palm as photosynthetic starch are actively being stored in the trunk, a mechanism that is truly one-of-its-kind among other starch producing crops. Furthermore, there have been research focusing on proteomic comparison between trunking and non-trunking sago palm where it is useful to differentiate stunted individuals (devoid of starch in the trunk) (Hussain et al., 2012; Hussain et al., 2020). Several proteins were found to be expressed differentially (Hussain et al., 2012) and the discovery of the differences in organellar copy number between trunking and non-trunking sago palm would further improve the understanding of this phenotype and reduce economic loss in

the industry in future. Therefore, in this chapter, the organellar genomes copy number at different organs, growth stages, localities and phenotypes of the sago palm were investigated to further improve our elucidation of organellar genome variations under the selected circumstances.

3.2 Methodology

3.2.1 Sampling and DNA Extraction

The living germplasm collection area in the east campus of Universiti Malaysia Sarawak (1°27'52.3"N, 110°27'04.7"E) is the initial collection site for a total of six sago palm organs such as aged leaf, pneumatophore, mature root, mature leaf, young root and young sucker shoot, each tissue sample was harvested in triplicates. Mature leaf tissue samples were collected from sago palm trees at five major developmental stages (*Plawei*, *Plawei Mani*t, *Bubul*, *Angau Muda* and *Angau Tua*) planted at sago palm plantation site in Mukah, Sarawak, Malaysia. Subsequently, the sampling of mature leaf from *Plawei Mani*t staged trunking sago palms was conducted at four main Sarawak localities, namely Pusa, Oya-Dalat, Samarahan and Mukah (trees from Pusa were transplanted in the living germplasm collection area of Universiti Malaysia Sarawak east campus). A sum of three leaf blades from three different leaf fronds was harvested from each tree and was later pooled for DNA extraction; three biological replicates were included to represent each organ, localities, phenotypes and growth stages tested. Prior to DNA extraction, 1 g of tissue sample was grounded to powder form using liquid nitrogen. The DNA extraction was performed emulating the manufacturer's protocol using Plant/Fungi DNA isolation kit (Norgen, Canada). The DeNovix DS-11+ spectrophotometer (DeNovix, USA) was

employed for the quality check of the extracted DNA in terms of optical density at 260/280 ratio and concentration.

3.2.2 Long-PCR

The mitogenome and plastome were represented by *nad4* and *rps14* genes in the long-PCR assay as both genes are single copy genes in the respective organellar genomes they embody. Two pairs of long-PCR primers [*nad4*_LF1 (5'-ATGGTCTCTTCAAGCCCCTC-3'); *nad4*_LR1 (5'-GTGAACACCCATCCGAACA-3'); *rps14*_LF1 (5'-TTCAGAGGGAGAAGAAGCGG-3'); *rps14*_LR1 (5'-TGATTCCTTCGGACGTGTCA-3')] were designed according to the conserved regions obtained from the multiple alignments of closely related species, namely oil, coconut and date palm (date and coconut palms mitogenomes with GenBank accession numbers JN375330.1 and KX028885.1 respectively as well as date and oil palms plastomes with GenBank accession numbers GU811709.2 and JF274081.1 respectively). These primer pairs targeted a ~10 kb fragment of mitogenome and plastome respectively. The 20 µL long-PCR reaction mixture consisted of 2.5 units KOD FX Neo high fidelity *Taq* DNA polymerase (Toyobo, USA), 2X KOD FX Neo PCR buffer (Toyobo, USA), 200 µM forward and reverse primers, 400 µM dNTP mix and 0.5 ng total genomic DNA. The long-PCR thermal cycling parameters were as follow: 1 cycle of 95 °C for 3 min; 22 cycles of 95 °C for 30 s, annealing at 64.8 °C (for ptDNA) or 54.4 °C (for mtDNA) for 1 min, 72 °C for 7 min; 1 cycle of final extension at 72 °C for 10 min. The long-PCR was conducted employing a T-100 thermal cycler (BioRad, USA). All reactions were done in triplicates with three biological and technical repeats.

The long-PCR amplicons were loaded in 4 µL volume into 1% agarose gel soaked in Tris-EDTA-acetate (TAE) buffer and the electrophorised gel was visualised under UV ray subsequent to the EtBr staining. A standard curve for DNA concentrations against EtBr-DNA fluorescence was constructed based on the size-fragmented GeneRuler 1 kb ladder (Fermentas, USA) on 1% agarose gel. This standard curve was employed to estimate the organellar genome count of all samples via ImageJ (Abramoff et al., 2004). The DNA copy number at the end of PCR was enumerated utilizing the formula “DNA copy number at the end of PCR = (Amount of DNA in ng X 6.022×10^{23})/(size of PCR product in bp X 10^9 X 650)”. The copies of organellar genome were calculated employing the formula “genome copy number = DNA copy number at the end of PCR/ $2^{\text{number of cycles}}$ ”. All organellar DNA copies from long-PCR were normalised to a single-copy nuclear DNA (*adh* copies obtained from qPCR). All organellar genome counts yielded via long-PCR were normalised to qPCR-determined organellar DNA copies (Kumar et al., 2014). Calculations were done for all replicates.

3.2.3 Real-time PCR

In this study, a total of three target genes *alcohol dehydrogenase (adh)* gene, *NADH dehydrogenase subunit 4 (nad4)* and *40S ribosomal protein S14 (rps14)* were chosen to represent the nuclear genome, mitogenome and plastome respectively. Three pairs of real-time qPCR primers targeting the three genes [adh_F1 (5'-AGTTTGATTACTCCGCGTGCA-3'), adh_R1 (5'-CCCACTTCATTGTAGCCCGC-3'); nad4_F1 (5'-GATGGGACTATGTGGTGAGGAA-3'), nad4_R1 (5'-AAAGAGCAAGAACCAAATCTCTCA-3'); rps14_F1 (5'-CGACGTTGTTTTTTGACTGGAAG-3'), rps14_R1 (5'-

GGTAACAAACATGCATGAACCAT-3') were designed based on conserved regions identified via multiple alignments across closely related species oil, coconut and date palm (date and coconut palms mitogenomes with GenBank accession numbers JN375330.1 and KX028885.1 respectively as well as date and oil palms plastomes with GenBank accession numbers GU811709.2 and JF274081.1 respectively). Each primer was designed to target a single copy gene in the respective genome, and this was cross-checked with BLASTN and Primer BLAST searches.

The qPCR was performed on an Applied Biosystems QuantStudio 5 real-time PCR system (ThermoFisher Scientific, USA) in a 10 µl reaction volume, employing 1 ng total extracted DNA as template, 0.3 µM forward and reverse primers, 5 µl ThunderBird SYBR Green qPCR master mix (Toyobo, USA). The thermal cycling parameters were set at 1 cycle of 94 °C for 3 min, followed by 40 cycles of 94 °C for 15 s, 60 °C for 15 s and 72 °C for 1 min. Melt curve analysis was also done at 95 °C for 15 s, 60 °C for 1 min and 95 °C for 1 sec (continuous acquisition). All reactions were performed in triplicates with three technical and biological repeats.

The minimum information for publication of quantitative real-time PCR experiments (MIQE) guideline established by Bustin et al. (2009) was inculcated in this study in the aspects of sample preparation, PCR efficiency, qPCR target information, DNA isolation, qPCR protocol and experimental design. All primer pairs have yielded PCR efficiency of 1 ± 0.1 and the standard curve was plotted utilizing cycling threshold values (Ct) yielded from tenfold dilution series. The fold difference of organellar genome count to the nuclear genome was enumerated and the genome counts of each sample tested were depicted in bar graphs. All reactions were done in triplicates.

3.2.4 Statistical Analysis

The relative organellar genome counts were subjected to one-way factorial analysis of variance (ANOVA) and Dunn's multiple comparison post-hoc test using PAST software (Hammer et al., 2001). The post-hoc test results were illustrated in bar graphs in the form of alphabets and different alphabets indicate significantly distinctive groups.

3.3 Results

3.3.1 Visualization of Long-PCR Bands

The intensities of the ~10 kb long-PCR mitogenome fragments (represented by *nad4* genes) of sago palm do not differ greatly when observed across all phenotypes, organs and growth stages prior to any normalization (Figure 3.1). On the contrary, the ~10 kb long-PCR plastome fragments displayed a myriad of intensities across all characteristics of sago palm tested. Across organs, the pneumatophore, aged leaf and mature leaf depicted a much intense band in comparison to that of mature and young roots. Across growth stages, the first three growth stages (*Plawei*, *Plawei Mani* and *Bubul*) portrayed some much higher intensities as compared to the final two growth stages (*Angau Muda* and *Angau Tua*). The intensity of *rps14* gene across of the trunking phenotype was observed to be much higher than that of spiny as well as non-trunking phenotypes where all of them were harvested from the *Plawei Mani* growth stage.

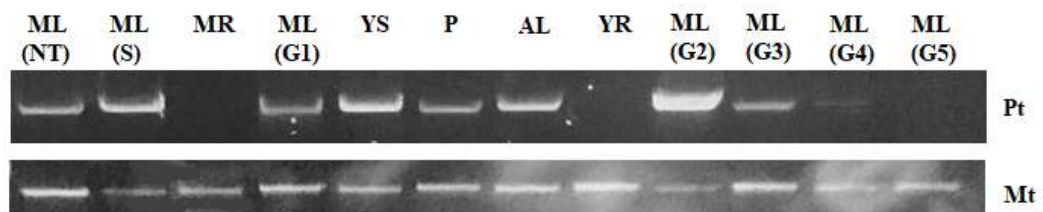


Figure 3.1: Long-PCR amplicons of mitogenome and plastome DNA across a variety of organs, growth stages and phenotypes of sago palm. Pt: plastome DNA; Mt: mitogenome DNA; ML: mature leaf; NT: non-trunking; S: spiny; YS: young shoot; P: pneumatophore; AL: aged leaf; YR: young root; G1: *Plawei* stage; G2: *Plawei Mani* stage; G3: *Bubul* stage; G4: *Angau Muda* stage; G5: *Angau Tua* stage.

3.3.2 Organellar Genome Analysis Across Different Organs

The organellar genomes across various organs of the sago palm along with their fold differences to nuclear genome were investigated in respect of the genome copy number, as depicted in Figure 3.2. The greatest plastome copy number was achieved by the mature leaf via the long-PCR method. The mature leaf scored the second place in terms of mitogenome copy number and shared the position with young root via the long-PCR approach. The fold differences recorded for mitogenome and plastome relative to that of the nuclear genome for mature leaf are 0.15- and 0.22-fold respectively at the mature leaf organ. The first position was taken by pneumatophore with plastome copy number recorded at over 5000. The organellar genome copy number for young root was high in regard to mitogenome (approximately 4000) but low in regard to plastome. A contradictory situation was detected in mature leaf organ.

The qPCR approach unravelled the peak amount of both plastome and mitogenome copy number in the mature leaf, which are 196780 and 188550 respectively (Figure 3.2). In other words, the plastome and mitogenome copy number of the mature leaf of sago palm was observed to be 7.5- and 7.2-fold relatively to the nuclear genome. Figure 3.2 showed the long-PCR/qPCR ratios of various organs in respect of mitogenomes and plastomes. In

mature leaf, the ratio was just about 0.25, which is half to that of the pneumatophore (close to 0.5). Mature leaf was selected for advanced analysis across different growth stages based on several factors as mentioned in the discussion section below.

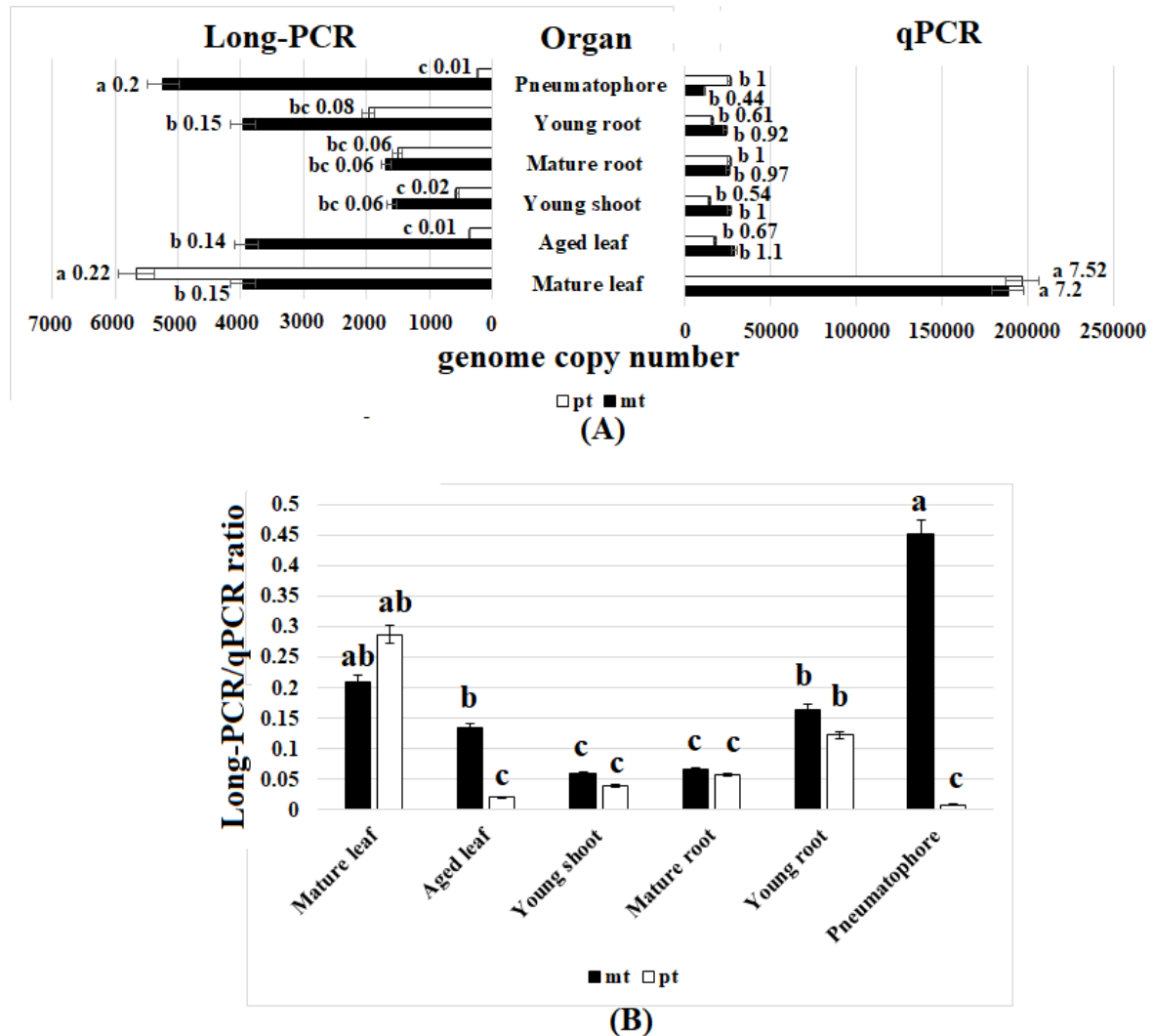


Figure 3.2: (A) The organellar genome copy number across different organs of sago palm established via long-PCR and qPCR. Statistical analysis was implemented using one-way ANOVA and post-hoc Dunn's Test with $p < 0.05$. Significantly distinct data are denoted with different alphabets and the fold differences to that of the nuclear genome copy number are depicted beside the alphabets. (B) The long-PCR/qPCR ratio graph displaying the plastome and mitogenome long-PCR/qPCR ratio across different organs of sago palm. mt: mitogenome; pt: plastome.

3.3.3 Organellar Genome Analysis Across Different Growth Stages

A total of five distinctive growth stages (*Plawei*, *Plawei Mani*, *Bubul*, *Angau Muda* and *Angau Tua*) were subjected to organellar genome copy number inspection (Figure 3.3). No significant differences in mitogenome copy number discovered across all growth stages tested with the exception of *Plawei* stage. The mitogenome fold differences of all growth stages were relatively high (0.18- to 0.19-fold) as compared to that of the *Plawei* stage. Besides, the *Plawei Mani* growth stage topped the list for the plastome copy number in the long-PCR reaction. The long-PCR data chart puts both *Angau Muda* and *Angau Tua* growth stages almost on par (0.09- and 0.11-fold) in respect of the mitogenome fold differences and the *Bubul* stage bottomed the list (0.01-fold).

The *Plawei Mani* stage once again outshined the other growth stages in terms of both plastome and mitogenome copy number determined via the qPCR method. The first runner-up is the *Plawei* stage with 5.83- and 6.51-fold differences observed across plastome and mitogenome respectively. Another notable difference detected between *Plawei* and *Plawei Mani* stage is that the plastome copy number is higher than that of mitogenome in *Plawei Mani* growth stage, but a contradiction was seen in *Plawei* stage. The other three growth stages were significantly different from the two abovementioned growth stages for both organellar genomes for the qPCR method. The long-PCR/qPCR ratio graph as displayed in Figure 3.3 yielded a uniform pattern across the final three growth stages with ratio close to 1 for mitogenome. For the plastome, the ratio increases from the first to final growth stage with a plunge observed in *Bubul* stage.

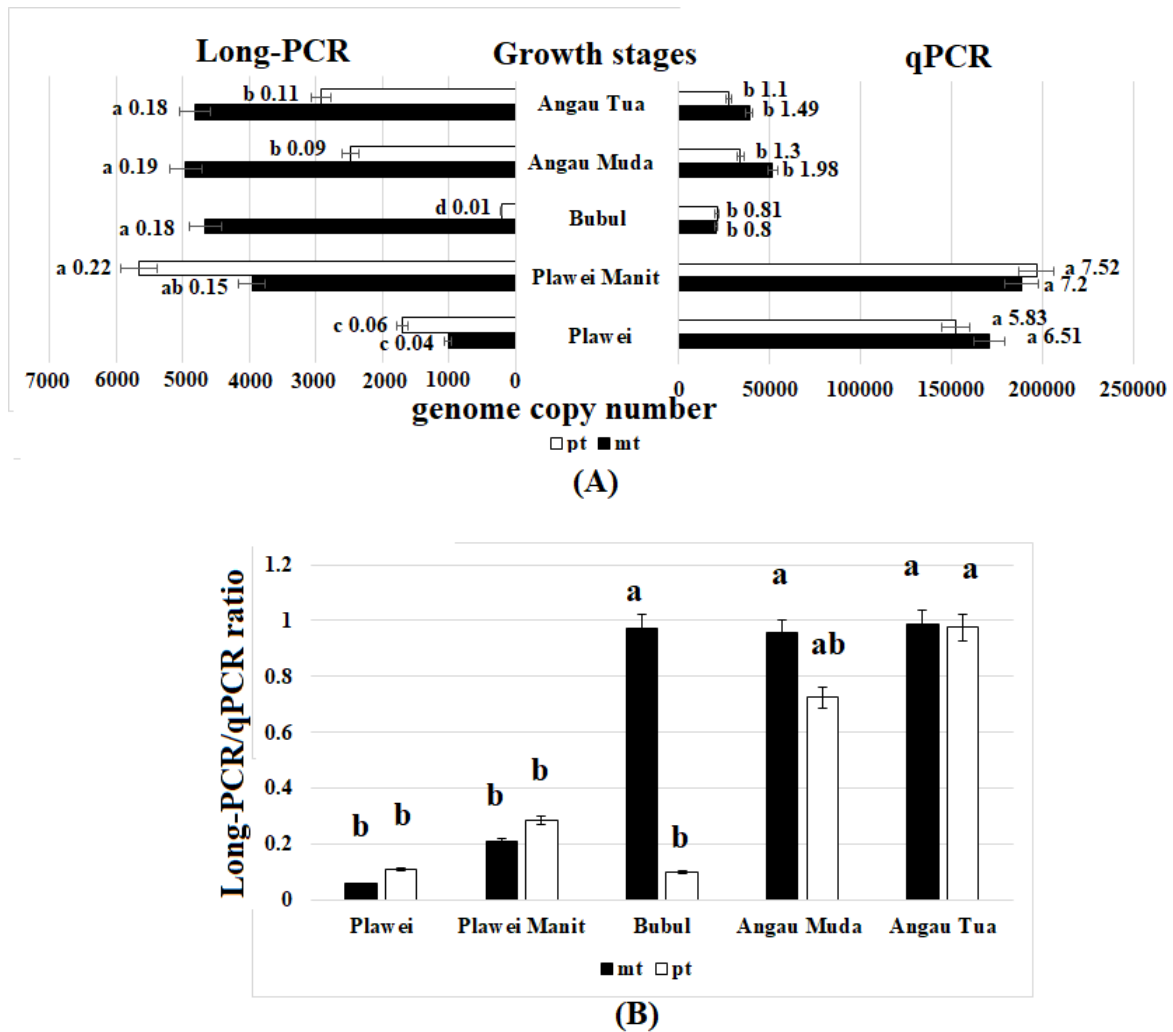


Figure 3.3: (A) The organellar genome copy number across different growth stages of sago palm established via long-PCR and qPCR. Statistical analysis was implemented using one-way ANOVA and post-hoc Dunn's Test with $p < 0.05$. Significantly distinct data are denoted with different alphabets and the fold differences to that of the nuclear genome copy number are depicted beside the alphabets. (B) The long-PCR/qPCR ratio graph displaying the plastome and mitogenome long-PCR/qPCR ratio across different growth stages of sago palm. mt: mitogenome; pt: plastome.

3.3.4 Organellar Genome Analysis Across Different Phenotypes

The three phenotypes tested for their organellar genome profiles in this study are spiny, trunking and non-trunking where they are all harvested from *Plawei Mani* growth stage (Figure 3.4). In the long-PCR reaction, the trunking phenotype scored the highest

among the three for both organellar genomes investigated. For instance, the mitogenome and plastome fold differences for spiny phenotype are 2.6 times and seven times lower than that of the trunking phenotype. For the non-trunking phenotype, the trend in the mitogenome and plastome fold differences is contradictory to that of the spiny phenotype but at similar degree of differences (7- and 3-fold differences respectively).

The qPCR approach yielded another interesting data in regard to the superior quality of the trunking phenotype. Figure 3.4 illustrated the inferior mitogenome and plastome fold difference amounts of both spiny and non-trunking phenotypes. The degree of fold differences of spiny sago palm to that of the trunking sago palm is around seven times for both organellar genomes. Similarly, both the mitogenome and plastome genome fold differences of the non-trunking sago palm is approximately seven times lower than its trunking counterpart. Looking at Figure 3.4, the long-PCR/qPCR ratios of all three phenotypes of sago palm showed no specific pattern. The highest ratio was detected in plastome of non-trunking sago palm whereas the lowest ratio belongs to that of plastome spiny sago palm.

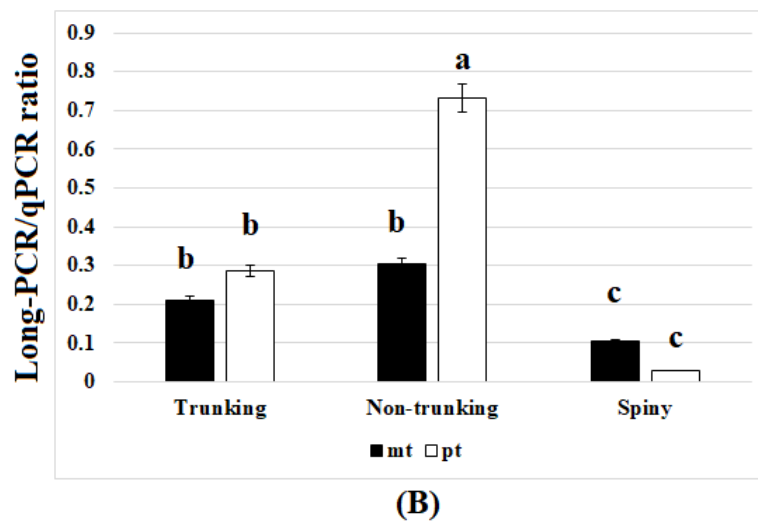
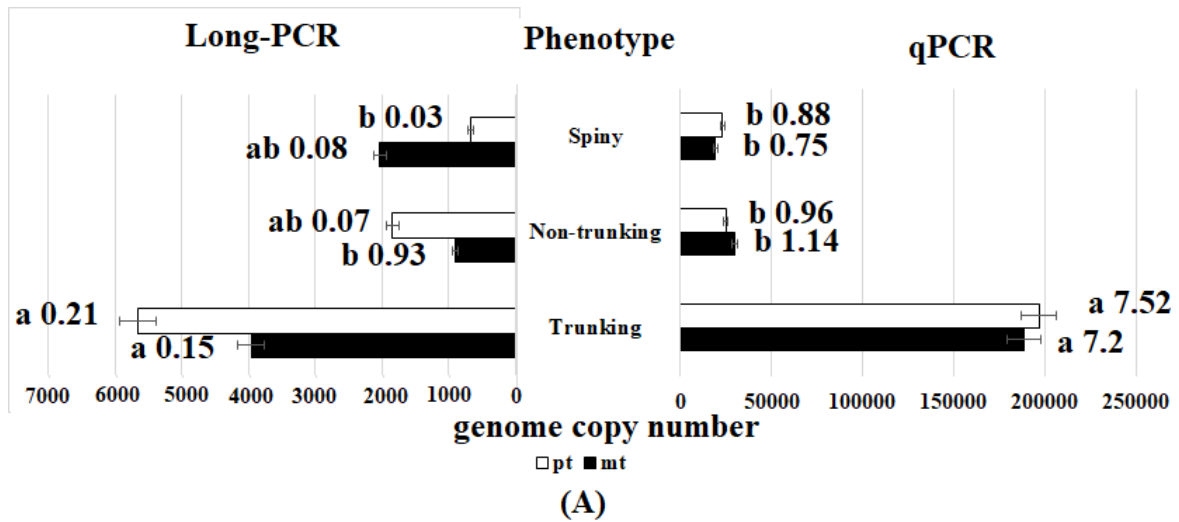


Figure 3.4: (A) The organellar genome copy number across different phenotypes of sago palm established via long-PCR and qPCR. Statistical analysis was implemented using one-way ANOVA and post-hoc Dunn's Test with $p < 0.05$. Significantly distinct data are denoted with different alphabets and the fold differences to that of the nuclear genome copy number are depicted beside the alphabets. (B) The long-PCR/qPCR ratio graph displaying the plastome and mitogenome long-PCR/qPCR ratio across different phenotypes of sago palm. mt: mitogenome; pt: plastome.

3.3.5 Organellar Genome Analysis Across Different Localities in Sarawak

The organellar genome copy numbers of sago palm were also examined across four main localities in Sarawak, Malaysia, namely: Pusa, Oya-Dalat, Mukah and Samarahan

(Figure 3.5). All samples from three different phenotypes were harvested during the *Plawei Manit* growth stage at the abovementioned localities (with the exception of Pusa and Samarahan where there is a lack of availability of non-trunking phenotype). At a glance, the organellar genome fold differences do not show any variations in the aspect of localities in Sarawak, Malaysia, even across mitogenome and plastome. Subsequent statistical analyses had further verified the raw observation that there is no significant difference across these organellar genome copy numbers across Sarawak main localities. For all localities, the trunking sago palm is superior in terms of both mitogenome and plastome fold differences against spiny and non-trunking phenotypes. For long-PCR approach, the non-trunking sago palm transcended its spiny counterpart with reference to plastome copy number and the vice versa occurred in terms of mitogenome. For qPCR method, both non-trunking and spiny phenotypes displayed similar levels of organellar genome fold differences.

Based on Figure 3.5, the long-PCR/qPCR ratio graph displayed a trend similar to that determined previously in Figure 3.4 for the trunking phenotype, for all phenotypes. The long-PCR/qPCR ratio for the spiny sago palm is the highest (close to 1) for plastome, especially sago palm from Oya-Dalat, followed by Mukah, Samarahan and Pusa. The long-PCR/qPCR ratios are most uniform across both organellar genomes for the trunking phenotypes for all four Sarawak localities. For the mitogenome, the long-PCR/qPCR ratio for the non-trunking phenotype outshined the other two phenotypes, achieving a ratio close to 0.8 for sago palm from Mukah and Oya-Dalat localities. The low long-PCR/qPCR ratios were observed in plastomes of trunking and non-trunking sago palm as well as mitogenomes of trunking and spiny sago palms.

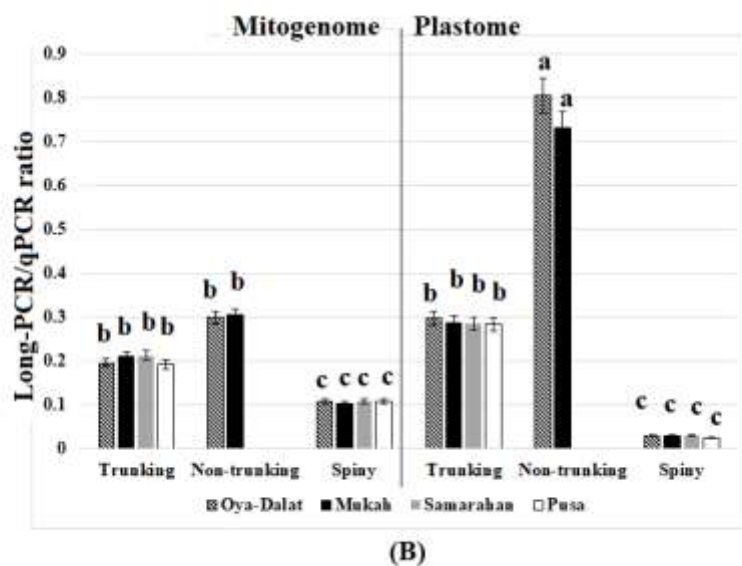
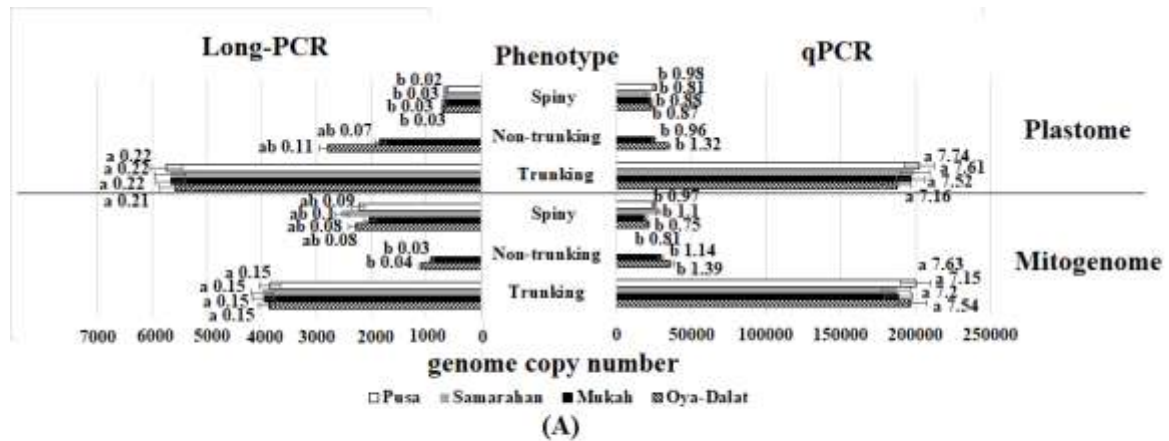


Figure 3.5: (A) The organellar genome copy number across different main localities of sago palm in Sarawak, Malaysia established via long-PCR and qPCR. Statistical analysis was implemented using one-way ANOVA and post-hoc Dunn's Test with $p < 0.05$. Significantly different data are conveyed with different alphabets and the fold differences to that of the nuclear genome copy number are indicated beside the alphabets. Non-trunking phenotype samples were not harvested from Pusa and Samarahan due to unobtainability. (B) The long-PCR/qPCR ratio graph illustrating the mitogenome and plastome long-PCR/qPCR ratio across different main localities of sago palm in Sarawak, Malaysia. Non-trunking phenotype samples were not harvested from Pusa and Samarahan due to unobtainability. mt: mitogenome; pt: plastome. Bars denote Sarawak localities.

3.4 Discussion

The long-PCR method is one effectual assay to identify the amount and presence of DNA damage (represented by the amplified 3 to 20 kb single copy gene from the genome) where DNA lesions can halt *Taq* DNA polymerase actions. In 2014, Kumar et al. (2014) introduced a term "DNA impediments" to accurately provide definition to any attribute that impede PCR amplification and to exclude all other causes of DNA lesions like single-strand gaps and double-strand lesions. On the other hand, the qPCR assay is another effective approach for the sensitive amplification of 100 to 200 bp DNA fragment from the genome regardless of size and molecular integrity. The qPCR approach often leads to the overestimation of genome copy number due to its high sensitivity as well as more than one targeted fragment may come from a single copy gene of the genome (Kumar et al., 2014). On the contrary, the long-PCR provides higher accuracy in describing the exact genome copy number, but the downside is that underestimation may occur as a result of physiological needs and developmental changes (Kumar et al., 2014). The combinatory approach that involves both qPCR and long-PCR is one powerful solution to capture both impeded and unimpeded DNA of the genome, and further provides us a more comprehensive view of the actual genome copy number. Lower long-PCR/qPCR ratio indicates lower levels of unimpeded DNA in the genome, and a ratio close to one represents a balance in the amounts of unimpeded and impeded DNA within the genome.

3.4.1 Organellar Genome Analysis Across Different Organs

The comparison of organellar genome across six plant organs (namely pneumatophore, young root, mature root, young shoot, aged leaf and mature leaf) in terms of copy number and fold difference was demonstrated in Figure 3.2. One of the highlights

observed is that the mature leaf scored the highest in the aspect of plastome genome copy and second to pneumatophore in respect to plastome copy number under the long-PCR approach. The relatively high 0.15- and 0.22-fold extra mitogenome and plastome copies to the nuclear genome in mature leaf elucidate the essential role of mature leaf in photosynthesis. The highest plastome copy number (over 5000) was detected in pneumatophore, the abundant plastome is required to cater its function as active oxygen absorber from air directly whenever there is a dissolved oxygen scarcity at the root. The young root is high in mitogenome content than its plastome content because active supply of ATP is the major requirement for the elongation of the young root. On the other hand, a contradictory phenomenon was discovered in mature leaf where it necessitates plastome more than mitogenome, determined via the long-PCR assay. These findings were in line with that conducted on maize seedlings (Kumar et al., 2014; Ma & Li, 2015).

From the qPCR point of view, the mature leaf was significantly superior to all the other organs tested in respect to plastome and mitogenome copy numbers. At least 7 times differences were revealed in mature leaf in contrast to other organs, emphasizing its major role played in dynamic photosynthesis to fulfil its physiological and developmental requirements. The highest long-PCR/qPCR ratio recorded among all six organs is the mitogenome of pneumatophore. A ratio close to 0.5 scored by the pneumatophore is a clear indication that half of the qPCR-determined mitogenome fragments are intact and functional, creating a big contrast to its plastome score (Figure 3.2). The mature leaf has around a quarter of its qPCR-determined organellar genome copy number intact as well as unimpeded, and this ratio is supported by the high long-PCR-determined functional genome copy number. The high impeded DNA of mature leaf as compared to that of pneumatophore is probably due to the mature leaf being constantly exposed to ultraviolet

rays and reactive oxygen species (ROS) (Boesch et al., 2010; Alexeev et al., 2013). Therefore, the mature leaf is selected for further analysis as it yields high value for easy detection and it hosts the most active photosynthesis process across various growth stages.

3.4.2 Organellar Genome Analysis Across Different Growth Stages

The organellar genome copy number and fold difference were investigated across five distinctive growth stages, encompassing *Plawei*, *Plawei Mani*, *Bubul*, *Angau Muda* and *Angau Tua* (Figure 3.3). The long-PCR mitogenome copy numbers across all four growth stages do not show any significant differences, only the *Plawei* growth stage is significantly different from the rest. The possible explanation is that the *Plawei* growth stage involves maximum vegetation growth where the trunking process is not induced yet. All other growth stages except *Plawei* are involved in high energy consumption processes such as inflorescence emergence, inflorescence development, flowering and fruiting.

The qPCR-determined organellar genome copy numbers of the *Plawei Mani* growth stage were the highest among other growth stages (Figure 3.3). On the side note, this growth stage also scored the highest long-PCR plastome count (5658). These data show the drastic energy requirement transition from *Plawei* to this growth stage, which correlate with the increasing organellar genome counts. Similarly, the notable plummet in qPCR plastome count at *Bubul* stage is probably another marker of energy transition from inflorescence emergence to inflorescence development. The demand for plastid synthesis (for instance, amyloplasts) elevates after *Bubul* stage, compacting the trunk with photosynthetic starch and this explains for the increasing plastome counts (Ogita et al., 1996). Close-to-1 ratios were observed across the mitogenomes of the final three growth stages of sago palm (Figure 3.3), it can be deduced that functional mitogenome counts

were maintained from the inflorescence development stage up to the fruiting stage. Despite experiencing fragmentation of single copy organellar genome genes (low long-PCR/qPCR ratio) probably due to numerous ROS attacks, the *Plawei Mani* growth stage yielded the highest mitogenome and plastome counts. Thus, this growth stage is the most feasible for further analysis across different phenotypes because most of the active biological processes occur during this stage.

3.4.3 Organellar Genome Analysis Across Different Phenotypes

For the organellar genome analysis across different phenotypes, a sum of three sago palm phenotypes, spiny, non-trunking and trunking, were investigated subsequent to harvesting from the *Plawei Mani* growth stage (Figure 3.4). The superior phenotype of sago palm determined from this study is that of the trunking phenotype with regards to organellar genome counts for both qPCR and long-PCR assays. In qPCR reaction, the trunking phenotype outshined the other two phenotypes by seven to ten times for both organellar genome whereas the same phenotype transcended the other two by three to seven times in the long-PCR method for both organellar genomes. There is no significant difference discovered across spiny and non-trunking phenotypes in both assays but interestingly the mitogenome and plastome counts displayed a vice versa relationship across these two phenotypes.

Furthermore, Rauwerdink et al. (1986) and Ehara et al. (2018) had verified that genetic distance does not correlate with spine formation in sago palm. To the best of our knowledge, the link between organellar genomes and spine development remains unearthed to date. The reason why the spiny and non-trunking sago palms required miniature amount of organellar genome count is that spines are effective in manifestation

reduction of natural pests like sago beetle (*Rhynchophorus vulneratus*) larvae and there is no food source for these larvae in non-trunking sago palm as they are trunk borers (Kimura, 1979; War et al., 2012). Therefore, the trunking sago palm probably necessitates extra organellar genomes to secrete secondary metabolites to further reduce the sago beetle attacks as well as to heal from the bore wounds. Furthermore, most of the plastome (70%) in non-trunking sago palms were intact and functional based on the long-PCR/qPCR ratio in Figure 3.4. Its incapability in trunk formation has most likely leads to its elevation of functional intact plastome counts in order to minimise the unfavorable photoinhibition effects and fulfill its photosynthetic requirements (Rijkers et al., 2001).

3.4.4 Organellar Genome Analysis Across Different Localities in Sarawak

A total of four main sago palm plantations at four different Sarawak localities (Samarahan, Pusa, Mukah and Oya-Dalat) were selected as harvesting sites for this study (Figure 3.5). All three phenotypes of sago palm were harvested during the *Plawei Mani* growth stage (non-trunking sago palms from Pusa and Samarahan were not sampled due to sample unavailability). Across all four Sarawak localities tested, no significant difference was observed for both organellar genomes investigated. Looking at the long-PCR/qPCR ratios (Figure 3.5), a similar phenomenon was revealed where these ratios are not distinctive from one another significantly among the four localities in Sarawak, Malaysia. These results from the present study are in agreement with various previous literatures stating that only one species of sago palm is recognised across countries like Thailand, Papua New Guinea, the Philippines and Indonesia determined via ITS markers, RAPD, AFLP and chloroplast markers approaches (Rauwerdink, 1986; Kjær et al., 2004; Ehara, 2009; Abbas et al., 2010; Enguito & Nevero, 2018; Abbas et al., 2020).

3.4.5 Interspecies Comparison of Organellar Genome Counts

There are not many plant species being subjected to organellar genome analysis in terms of genome copy number and fold difference. Most of them are important food crops like maize, beet and wheat where they play pivotal roles as high economic commodities. One of the most studied food crops in this field is the maize plant. The highest unimpeded mitogenome fold difference determined in this study is 0.2-fold, which was yielded from pneumatophore of the trunking sago palm. The mitogenome fold difference range in maize seedlings fell within 0.1 to 3.2 (Kumar et al., 2014) whereas the sago palm mitogenome fold range was slightly lower (in the range of 0.01 to 0.2). In higher plants, the organellar genome counts tend to be insignificant to that of their nuclear counterparts, particularly throughout leaf senescence and development whereas smaller plants need extra copies of organellar genomes to buffer DNA damages induced from their external environment (Li et al., 2006).

On the other hand, the highest unimpeded plastome count reported in this study is 5658 (0.22-fold to the nuclear genome) which this was obtained from mature leaf of *Plawei Mani* staged trunking sago palm. The plastome count of light-cultivated wheat was between 800 to 1000 per plastid as developmental progresses albeit the increment in chloroplast DNA and plastid count with cell age (Boffey & Leech, 1982). No significant differences observed across young to old rosette leaves of *Arabidopsis* but the plastome counts fluctuates between 1000 to 1700 per cell (Zoschke et al., 2007). For maize seedlings, the plastome copy number was 2300 (Udy et al., 2012) and 1498 (Ma & Li, 2015) in less mature lower part of the leaf. In beet, the plastome counts fell within the range of 1100 to 1900 per cell (Tymms et al., 1983). In short, this chapter has documented the organellar genome copy changes that are useful for future gene expression studies.

CHAPTER 4

GENOME SURVEY ANALYSIS

4.1 Introduction

The nuclear genome orchestrates almost every functional regulation vital for the survivability and adaptability of the host plant. The plant nuclear genome can encompass up to 120 000 Mbp of nucleotides as found in lilies (one of the largest plant genomes), whereas on the other hand it can come in miniature form housing as little as 125 Mbp nucleotides as discovered in roses, crucifers, pteridophytes and bryophytes (Heslop-Harrison, 2003). The organellar genomes may have their own gene regulation controls but the master control is still in the hands of the nuclear genome (Gatehouse, 1997).

The plant nuclear genome exerts control on gene regulation of both organellar genomes through post-transcriptional control as well as nuclear factors (Gatehouse, 1997). One notable phenotypic impact of the plant nuclear genome is the distinctive morphologies of plastid among plant organs. For instance, the light capturing plastids are enriched within the leaf organ whereas these plastids are absent in the storage organs. Interestingly, the nuclear genome has the capability to remove several functional genes from the organellar genomes and include them back into the nuclear genome via the swap of genetic contents across these genomes (Gatehouse, 1997).

The *Arabidopsis thaliana* nuclear genome is one of the earliest higher plant genomes sequenced since over two decades ago (Initiative, 2000). Since then, the exponential increment in the plant genome database has flourished with the aid of the booming next generation sequencing technology advancement (Goff et al., 2002; Paterson

et al., 2009; Schnable et al., 2009). The sago palm is one of the most underwhelming candidates that requires plenty of molecular research attention due to its exclusive economic status (Nisar & Hussain, 2010). The mystery behind its high salinity tolerance and starch yield can only be resolved with the availability of the whole nuclear genome sequences. Hence, in this chapter, a low-coverage genome survey was performed on the sago palm to unravel several key characteristics of the genome such as GC content, repeated sequences, RNA cistrons as well as heterozygosity.

4.2 Methodology

4.2.1 Sampling and DNA extraction

The sago palm leaf tissues were harvested from the same tree utilised for the chloroplast genome sequencing following authentication by the plant specialist working in the living germplasm of Universiti Malaysia Sarawak East Campus (1°27'52.3"N, 110°27'04.7"E). The DNA extraction was done as previously described, utilizing Plant/Fungi DNA isolation kit (Norgan Biotek, Canada). Quality check (in terms of optical density 260/280 ratio as well as DNA concentration) of the DNA extracted was executed using DeNovix DS-11+ spectrophotometer (DeNovix, USA). Agarose gel electrophoresis was then employed to verify the integrity of the template.

4.2.2 Sequencing and Quality Control

The sago palm DNA was sheared using a Covaris Ultrasonicator (Covaris, United Kingdom) into 350 bp fragments. The sheared DNA was consequently handled using NEB Ultra II library preparation kit (NEB, United Kingdom) according to the manufacturer's

guidelines. The library was successively enumerated and sequenced on an Illumina HiSeq X platform located at the Beijing Genome Institute. The raw pair-end reads were later adapter- and quality-trimmed using fastp v. 0.18 (Chen et al., 2018) to be used for ensuing analysis.

4.2.3 K-mer Analysis

The frequency of the trimmed paired-end reads as well as the approximation of the k-mer richness (k-mer size of 21, 25 and 31) were conducted utilizing KMC 3 (Kokot et al., 2017). The k-mer histogram produced for each k-mer size was sent to the GenomeScope webserver (Vurture et al., 2017) to approximate the sago palm repeated sequences, genome size and heterozygosity. GenomeScope parameters were as set at max K-mer coverage 1000 as well as read length 150 bp.

4.2.4 *De novo* Genome Assembly and Genome Completeness Assessment

The sago palm genome contigs and scaffolds were constructed using Platanus (Kajitani et al., 2014). The clean reads were first trimmed before the contig generation procedure, applying only trimmed paired-end reads. Next, the sago palm contigs were subjected to scaffolding utilizing trimmed mate-pairs. Both types of reads were applied to fill the gaps through the gap-closing method to achieve the sago palm genome assembly. The genome comprehensiveness of the sago palm was examined employing BUSCO (Simão et al., 2015). A sum of 1374 single copy plant genes were utilised to examine the genome completeness of the sago palm, and at the same time the percentage of complete, fragmented and missing genes were counted.

4.2.5 Microsatellite Analysis

The microsatellite analysis was executed applying Kmer-SSR (Pickett et al., 2017) on the handled reads of the sago palm genome. The types of microsatellites involved in the analysis are dinucleotides, trinucleotides, tetranucleotides and pentanucleotides. Only microsatellites with at least four repeats were chosen for analysis.

4.2.6 Starch-related Homology Search

The genome scaffolds of the sago palm were analysed using protein BLASTx via DIAMOND (Buchfink et al., 2015) for homology search. The reference protein sequences were downloaded from GenBank database in reference to that of Ohdan et al. (2005) on rice (*Oryza sativa*). A sum of 14 starch-related protein sequences was chosen to form the reference list (Table 4.1). Leading identity scoring sago palm genome scaffolds with a cut-off value larger than 70% and the aligned length greater than 100 aa were nominated for further analysis. The consequential result was envisioned through graph plots and tabulation.

Table 4.1: List of protein sequences utilised for DIAMOND protein BLASTx with their respective GenBank accession numbers.

Protein	GenBank accession number
Branching enzyme-1 precursor	BAA01855.1
Branching enzyme-3 precursor	BAA03738.1
Starch branching enzyme rbe4	BAA82828.1
Isoamylase	BAC75533.1
Putative soluble starch synthase IV 1	AAM49812.1
Starch synthase IVb precursor	AAQ82623.1
Soluble starch synthase II 1	AAK64284.1
Soluble starch synthase II 2	AAK81729.1
Granule binding starch synthase II precursor	AAL58572.1
Soluble starch synthase precursor	BAA03739.1
Putative soluble starch synthase II 3	AAL16661.1
Putative soluble starch synthase III	AAL40942.1
Putative soluble starch synthase III 2	AAM49811.1
Starch (bacterial glycogen) synthase	CAA44065.1

4.2.7 Identification of rRNA Cistron Components and Phylogenetic Analysis

The rRNA cistron elements were discovered through NOVOPlasty (Dierckxsens et al., 2017). The rRNA cistron components (5.8S, ITS1, 18S, ITS2 and 28S) sequences were uncovered from the constructed contigs of the sago palm genome. The ITS1 and ITS2 regions were uncovered from the 18S contig via ITSx (Bengtsson-Palme et al., 2013). The ITS2 sequences of 119 palm species and two gymnosperm outgroups (*Pinus contorta* and *Ginkgo biloba*) were downloaded from the public GenBank database for phylogenetic analysis. A model test was performed before the tree construction and the most suitable model tested is the Kimura 2 + G model. The maximum likelihood phylogenetic tree was constructed applying MEGA X (Kumar et al., 2018) accordingly with bootstrap replications of 1000.

4.3 Results

4.3.1 Sequencing Data Statistics

A total of 124,258,713,300 paired-end reads were yielded summing up to 124.26 Gb. After the low-quality reads were eliminated through the stringent screening process, a sum of 123,430,321,878 clean bases were harvested from the raw bases (Table 4.2). The predicted sago palm genome size was 464,306,862 to 616,599,600 bp with GC content of 37.31%, genome heterozygosity of 0.63%-0.66%, as well as genome repetitiveness of 31.9%-35.7%. The BUSCO completeness of scaffolds and contigs were very similar in terms of the percentage of single-copy, duplicated, fragmented and missing genes. The amount of both scaffolds and contigs do not differ much, same goes to the N50 values.

Table 4.2: The summary of general genomic information of sago palm nuclear genome.

Organism	<i>Metroxylon sagu</i>
Bioproject	PRJNA6606601
Biosample	SAMN15953627
Accession number	JACVBY0000000000
Scaffold BUSCO completeness	21.5% single-copy, 1.1% duplicated, 32.2% fragmented, 45.2% missing
Contig BUSCO completeness	21.3% single-copy, 1.1% duplicated, 31.9% fragmented, 45.7% missing
Scaffold N50	710 bp
Contig N50	769 bp
Number of scaffolds	743,498,467
Number of contigs	739,583,200
Genome heterozygosity	0.63-0.66%
Genome repetitiveness	31.9-35.7%
Predicted genome size	464,306,862-616,599,600 bp
GC content (%)	37.31

4.3.2 K-mer Analysis

All clean reads of the sago palm genome were used for K-mer analysis employing 21, 25 and 31 K-mer values. K-mer frequency distribution graphs were drawn for every K-mer analysis conducted under this study (Figure 4.1). The y-axis portrayed the quantity of K-mers for the matching depth presented on the x-axis. According to Figure 4.1, the peaks produced for each K-mer triplicate are constant. The first peak depth (heterozygous peak) is 80 while the second peak depth (homozygous peak) is 159. The K-mer depth of the homozygous peak of the sago palm is in closely resembled that of the anticipated K-mer depth. The sago palm genome size was forecasted by the division of the K-mer number by the K-mer depth, the sago palm K-mer numbers obtained in this study range from 73,824,791,058 to 98,039,336,400. Therefore, the predicted genome size of the sago palm genome ranges from 464,306,862 to 616,599,600 bp. Moreover, the heterozygosity ratio was enumerated at 0.63% to 0.66%. The percentage of repeated sequences estimated at

31.9% to 35.7%, estimated as 1.8 times K-mer amount succeeding to the main peak over the entire K-mers quantity.

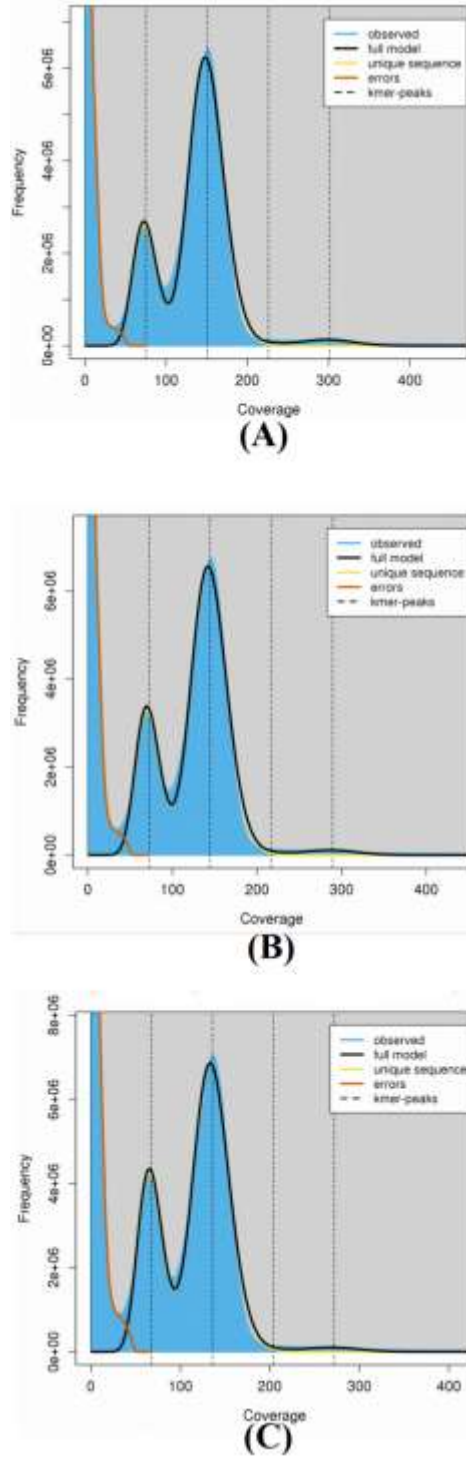


Figure 4.1: The distribution curves of K-mer. (A) K21: K-mer value 21; (B) K25: K-mer value 25; K31: (C) K-mer value 31. GenomeScope parameters: read length = 150 bp; max K-mer coverage = 1000.

4.3.3 GC Content and Distribution Pattern

The GC depth analysis unearthed a GC content coverage for most of the windows, in which they range from 30% to 50% (Figure 4.2). The sequencing depth was larger than 20-fold, where majority of them are positioned between 50 to 150-fold. There are no apparent irregularities detected among the sago palm samples aside from the nonappearance of GC bias. There are two layers discovered from the sago palm GC distribution graph and the bottom layer form a more reduced cluster than the top layer. The sequences of the low-depth distribution were scrutinised carefully, in accordance with that of Li et al. (2019). Contamination was not discovered subsequent to the authentication of these sequences against NCBI BLAST searches. The occurrence of the two layers is due to the heterozygosity landscape of the sago palm, whereby the reads within that area is precisely half of the whole genome product (Li et al., 2019).

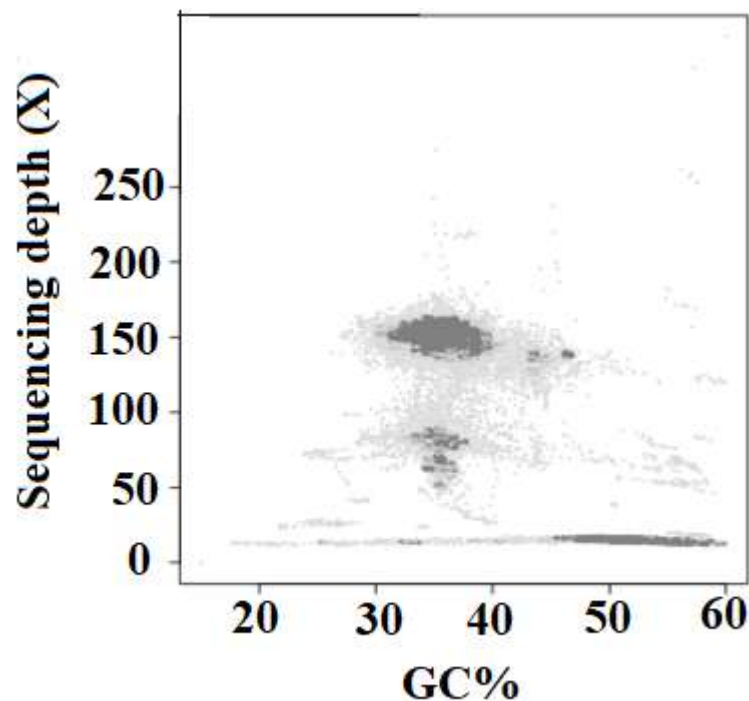


Figure 4.2: The GC depth graph of sago palm. X-axis: GC percentage; Y-axis: sequencing depth.

4.3.4 *De novo* Genome Assembly and Genome Comprehensiveness Assessment

The *de novo* sago palm genome assembly has generated 739,583,200 contigs as well as 743,498,467 scaffolds. The sago palm scaffold and N50 contig N50 were recorded at 710 bp and 769 bp respectively. The BUSCO analysis had estimated the scaffold comprehensiveness at 22.6% (1.1% duplicated and 21.5% single-copy) whereas the contig comprehensiveness was detected at 22.4% (1.1% duplicated and 21.3% single-copy).

4.3.5 Small Sequence Repeat (SSR) Analysis

Applying the Kmer-SSR script, a total of 14,929,016 SSRs enumerated from the entire sago palm genome. Majority of them are dinucleotides, encompassing a percentage of 62.24% (9,292,366 SSRs) (Figure 4.3). The tetranucleotides was the second most abundant population, covering 16.81% (2,509,158 SSRs) of the sum. Around 15.04% (2,244,728 SSRs) of them are found to be trinucleotide repeats. The residual 5.91% (882,764 SSRs) are the minority pentanucleotide repeats. The TC/GA, AT/TA and AG/CT are the leading dinucleotide SSRs where this trio took up 92.37% of the whole dinucleotide SSR group entities (Figure 4.4). For the trinucleotide repeats, nearly 50% the trinucleotide microsatellites are indifferent from TAT/ATA, AAT/ATT or TTA/TAA. The TTAT/ATAA, AAAT/ATTT and TTTA/TAAA SSRs ranked top on the tetranucleotide repeat frequency chart with 7.58%, 16.86% and 11.14% respectively, covering more than 25% of the entire tetranucleotide repeat group members. Likewise, 25% of the pentanucleotide SSR population is either TTTTA/TAAAA, AAAAT/ATTTT or AAAAG/CTTTT (6.12%, 11.98% and 8.27% respectively). On the other hand, the AT/TA SSR is the most abundant SSR found within the sago palm nuclear genome, with the frequency of 4,340,433 (29.07% of entire microsatellite population).

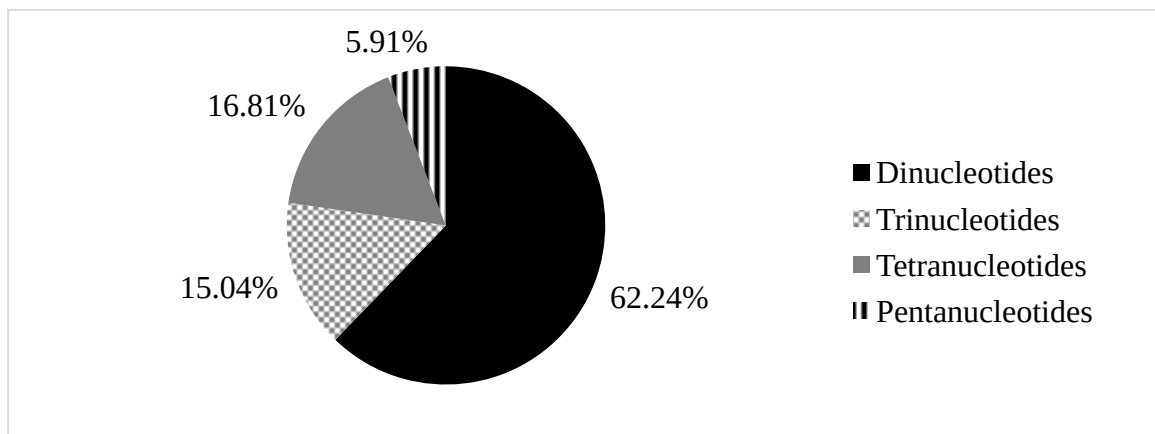


Figure 4.3: The constitution of dinucleotide, trinucleotide, tetranucleotide and pentanucleotide SSRs identified from the sago palm genome paired-end sequencing reads.

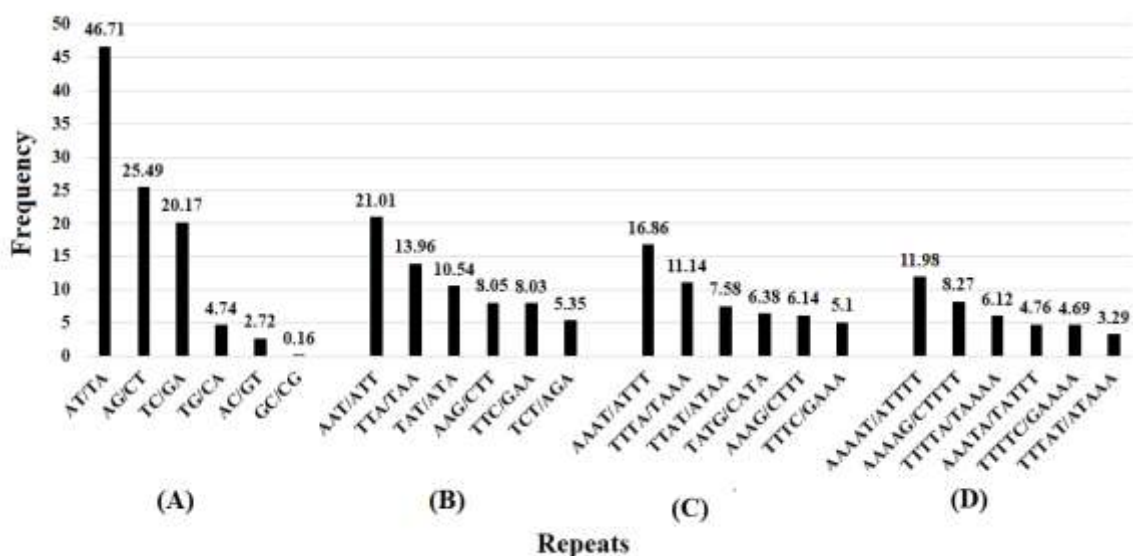


Figure 4.4: The SSR percentage bar graphs with the chosen top six frequency SSRs from each group: (A) dinucleotide repeats, (B) trinucleotide repeats, (C) tetranucleotide repeats and (D) pentanucleotide repeats.

4.3.6 Starch-related Homology Analysis

A sum of 238 sago palm genome scaffolds have been recognised as a direct presenter for starch-related routes and activities. These genome scaffolds have the genetic codes correlated with the starch breakdowns and passageways as characterised in rice

(*Oryza sativa*) by Ohdan et al. (2005). Most of the scaffolds (19.33%) are interrelated with branching enzyme-3 precursors as well as branching enzyme-1 (Figure 4.5). About 9.24% (22) of the discovered sago palm genome scaffolds are deemed to be participating in starch branching enzyme rbe4-like biological actions. Moreover, isoamylase are also plentifully scattered throughout these sago palm genome scaffolds with 8.40% constituent (20). Additional protein hits with fewer than 20 genome scaffolds are starch (bacterial glycogen) synthase, granule bound starch synthase II precursor, soluble starch synthases as well as soluble starch synthase precursor.

The leading identity scoring sago palm genome scaffolds with cut-off value larger than 70% and affiliated size lengthier than 100 aa were exhibited in Table 4.3. A total of nine sago palm genome scaffolds notched similarity percentages of 70% and above to that of the rice starch protein hits. Scaffold51695 scored the uppermost identity (87.4%) to that of branching enzyme-1 precursor, comprising of a 302 aa protein. Scaffold265298 recorded three protein hits that are all bettering the chart with hits from putative soluble starch synthase II 3, soluble starch synthase II 1 and soluble starch synthase II 2 (76.7%, 81.1% and 80.5% respectively). The longest protein hit is from Scaffold53770 with similarity index exactly at the 70% cut-off point. Other sago palm genome scaffolds have close resemblances with soluble starch synthases as well as branching enzyme precursors.

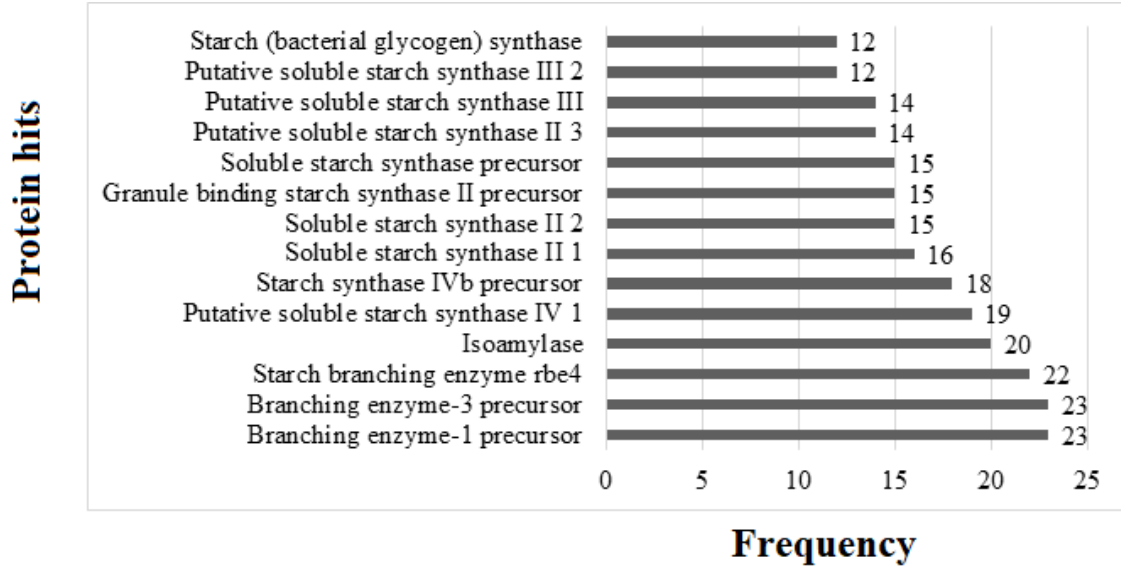


Figure 4.5: The genome-wide starch-related protein hits predicted utilising the sago palm genome scaffolds.

Table 4.3: Leading similarity scoring scaffolds (cut-off value 70%) with their aligned length greater than 100 aa, along with their respective GenBank accession number, E-value, identity percentage and amino acid length.

Scaffold	Protein hit	GenBank accession number	Identity (%)	Length (aa)	E-value
Scaffold51695	Branching enzyme-1 precursor	BAA01855.1	87.4	302	8.30 e ⁻¹⁶⁴
Scaffold265298	Soluble starch synthase II 1	AAK64284.1	81.1	159	7.80 e ⁻⁸⁰
Scaffold265298	Soluble starch synthase II 2	AAK81729.1	80.5	159	1.70 e ⁻⁷⁹
Scaffold265298	Putative soluble starch synthase II 3	AAL16661.1	76.7	159	5.80 e ⁻⁷⁵
Scaffold53770	Soluble starch synthase II 1	AAK64284.1	75.5	319	2.20 e ⁻¹⁵³
Scaffold476978	Branching enzyme-3 precursor	BAA03738.1	73.1	108	2.30 e ⁻⁴⁷
Scaffold53770	Soluble starch synthase II 2	AAK81729.1	72.7	319	2.60 e ⁻¹⁴⁶
Scaffold476978	Branching enzyme-1 precursor	BAA01855.1	71.2	111	3.10 e ⁻⁴⁴
Scaffold53770	Putative soluble starch synthase II 3	AAL16661.1	70	320	1.20 e ⁻¹³⁸

4.3.7 Characterisation of rRNA Cistrons and Phylogenetic Analysis

The 18S and 28S rRNA of the sago palm were unearthed from the genome contigs through NOVOPlasty (Dierckxsens et al., 2017). The length of 28S and 18S rRNA of sago palm is 4030 bp and 1810 bp respectively. Subsequent to the nucleotide BLAST against NCBI database, the best nucleotide hit comes from the coconut palm (*C. nucifera*) for both rRNA cistronic pieces, with 98% similarities. The e-value was documented at 0.0 for both nucleotide hits (Table 4.4). The ITS2 length of sago palm is 249 bp whereas the ITS1 is marginally longer, 264 bp. The 5.8S rRNA of the sago palm is recorded at 160 bp long.

Glancing at the phylogenetic tree (Figure 4.6), a sum of 13 tribes were identified, specifically Sabaleae, Arecoideae, Plectocomiinae, Coryphoideae, Eugeissoninae, Oncocalaminae, Pigafettinae, Metroxylinae, Ancistrophyllinae, Phoeniceae, Calaminae, Lepidocaryeae and Raphinae. The largest tribe is Arecoideae, encompassing 52.5% of all palms examined (63 members). The Arecoideae group is subdivided into two big bunches, flanking the Coryphoideae tribe. The sago palm inhabits in a mini constellation with its only congeneric entity *Metroxylon salomonense*, within the Metroxylinae tribe, with bootstrap value of 75%, which renders them a powerful clade. The other three members of the Metroxylinae tribe (*Korthalsia jala*, *K. cheb*, and *K. rogersii*) are positioned not too distant from the sago palm.

Table 4.4: Leading nucleotide hits of sago palm rRNA cistrons, alongside their similarity percentage, E-value and amino acid length.

rRNA	Nucleotide hit	Identity (%)	Length (bp)	E-value
18S	<i>Cocos nucifera</i> cultivar Chowghat Green Dwarf 18S ribosomal RNA gene, partial sequence	98	1810	0.0
28S	<i>Cocos nucifera</i> cultivar Chowghat Green Dwarf 28S ribosomal RNA gene, partial sequence	98	4030	0.0

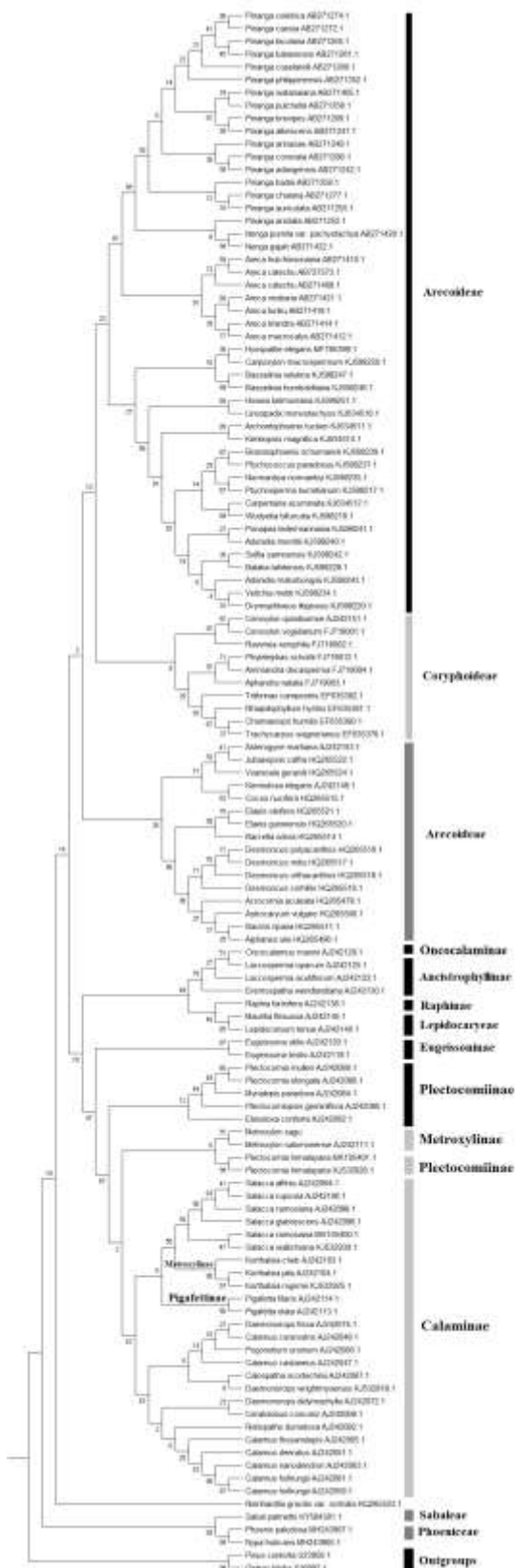


Figure 4.6: The ITS2 maximum likelihood phylogenetic tree of 120 palms (containing sago palm) with Kimura 2 + G model and 1000 bootstrap repetitions.

4.4 Discussion

4.4.1 Genome Length Approximation

Plant genome length data is indispensable in various studies involving adaptation, cell cycle examination as well as evolution (Shi et al., 2016). There is a myriad of methods to predict the genome length of an organism, varying from flow cytometry, pulsed-field gel electrophoresis, real-time PCR amplification, high-throughput sequencing, renaturation kinetics, to *in silico* K-mer prediction (Frenkel & Roizman, 1971; Chen et al., 1990; De Vita et al., 1997; Li et al., 2008; Chen et al., 2015; Roslan et al., 2020). The precision and duplicability of results are critical gauges of the dependability of the approaches employed. K-mer estimation is a very formidable instrument in approximating the genome length of unknown species and this approach has been demonstrated effective in plant species such as sword-leaf dogbane, monk fruit, Chinese tulip tree and camphor tree (Wu et al., 2014; Tang et al., 2015; Zhong et al., 2017; Li et al., 2019).

The genome length of the sago palm projected in this study is 464,306,862 to 616,599,600 bp. This genome length fell within the bottom range between the shortest plant genome 61 Mbp from *Genlisea tuberosa* and the lengthiest plant genome 150 Gbp from *Paris japonica* (canopy plant) documented thus far (Pellicer et al., 2010; Fleischmann et al., 2014). When contrasting the genome length of the sago palm across other palm family members, the sago palm genome size is thrice smaller than that of oil palm (1.8 Gbp) (Singh et al., 2013) but practically comparable to that of date palm (605.4 Mb) (Al-Mssallem et al., 2013). Remarkably, the sago palm genome has been formerly projected by Roslan et al. (2020) to be 1.87 Gbp by means of the real-time PCR method. The effort portrayed therein used the copy number of two housekeeping genes (elongation factor EF3 and beta-actin) to approximate the genome length with reference to that of the positive

control *Pichia pastoris*. Nonetheless, this approach of genome length projection might be imprecise owing to the occurrence of lofty errors (Gregory et al., 2013), causing in either underestimation or overestimation of the real genome length (Gregory et al., 2013). The irregularity of genome length approximation by the real-time PCR approach has been noticed in *Xiphophorus maculatus* where the projection was beyond 60% off (Wilhelm et al., 2003; Gregory et al., 2013). Furthermore, the genome length of *Musca domestica* was undervalued by no less than three-fold via the real-time PCR assay (Gao & Scott, 2006; Tsoumani & Mathiopoulos, 2012; Gregory et al., 2013).

4.4.2 Genome Heterozygosity, Repeated Sequences Percentage and GC Content

The GC content of the sago palm genome revealed in this study was 37.31%, which positions it within the typical array of 30% to 47% for most plants (Shangguan et al., 2013). This value is not far from that of *Picea abies* (37.8%), *Ananas comosus* (38.3%) and *Nelumbo nucifera* (38.2%) (Ming et al., 2013; Nystedt et al., 2013; Ming et al., 2015). However, the sago palm GC content is much greater in contrast to that of apple and alfafa (< 30%) but much lower in comparison to that of sorghum, maize and rice with GC content surpassing 40% (Li et al., 2019). Within the palm family, the GC content of the sago palm is greatly conserved across date palm (38.5%) and oil palm (37%) (Al-Dous et al., 2011; Singh et al., 2013). There will be more errors appear in the high-throughput sequencing whenever the GC content of the genome is at any one of the extreme spectrum edges (less than 25% and higher than 65%), interrupting the spliced data precision (Aird et al., 2011).

Both the heterozygosity ratio and the percentage of the repeated sequence hold necessary responsibilities in the splicing and fabrication of the genome (Li et al., 2019). In this study, the heterozygosity ratio of the sago palm was calculated as 0.63%, which

exactly echoed that of *Apocynum venetum* circumstantially (Li et al., 2019). This ratio in sago palm is much greater when contrasted among plants like Clementine, pummelo and Satsuma and (0.462%, 0.022% and 0.435% respectively) but much more diminished in contrast to that of sweet orange (0.716%) (Shimizu et al., 2016). Within the palm family, the heterozygosity ratio has not been documented in oil palm yet to date. The heterozygosity ratio of the date palm is discovered at 0.46% (Al-Dous et al., 2011), which is comparatively diminished when compared to that of the sago palm. The double peaks observed in the contig distribution graph (Figure 4.1) had verified the occurrence of heterozygosity in the sago palm genome. The sago palm heterozygosity values of 0.63% to 0.66% are considered as low heterozygosity values corresponding to the categorization described by Wu et al. (2014).

The repeated sequences percentage of the sago palm genome was computed as 31.9% to 35.7% from the K-mer distribution curves. A genome with this percentage smaller than 50% is considered as low repetitive genome (Wu et al., 2014). This percentage of sago palm is much smaller when contrasted among *Populus trichocarpa* (42%), *Vitis vinifera* (41.4%) and *Oryza sativa* (39.5%) (Al-Mssallem et al., 2013). Amongst the palm family members, the sago palm repeated sequences percentage is the smallest as the percentages in oil palm (57%) and date palm (38.41%) are much greater (Al-Mssallem et al., 2013; Singh et al., 2013). The expression and reassortment of the structural genes are orchestrated by these repeated sequences, apart from their indispensable contributions to the non-coding regions. As yet, our conception on the functions of these repeated sequences in sago palm is still sparse and additional inspection is critical in uncovering which repeated sequences hypotheses suited that of the sago palm, either adaptive theory, mutation-drift model or transposition mechanism (Shi et al., 2016).

4.4.3 Genome Comprehensiveness Assessment

The BUSCO genome comprehensiveness tests on both sago palm genome contigs and scaffolds had produced substantially low percentage of completeness. This implies that the sago palm genome from the short-read sequencing in this study is mainly missing and fragmented. This occurrence is possibly triggered by the highly repetitive nature of the sago palm genome. Hence, long-read sequencing is proposed to enhance the resolution of the genome (Mantere et al., 2019).

4.4.4 Genome-wide Microsatellite Analysis

Microsatellites are formidable genetic and molecular indicators that are crucial in researching codominant inheritance, multiple allele polymorphisms and species identification. In the current study, the GC content of the sago palm genome was 37.31%, which was much more diminished than the AT composition documented (62.69%). Glancing at the dinucleotide repeats, the AT/TA repeat is the most prolific repeat (found in 4,340,433 SSRs) amongst dinucleotides and amongst all microsatellites discovered within the sago palm genome, covering 29.07% of the entire sago palm genome microsatellite population. A comparable trend was reflected in the sword-leaf dogbane nuclear genome as well as the chloroplast genome of natangura palm (*Metroxylon warburgii*) and sago palm (Barrett et al., 2016; Li et al., 2019). In date and oil palm, the AG/CT notched the top with 85.7% and 56.4% respectively (Zhao et al., 2013; Kalyana et al., 2018). The AAT/ATT SSR is the most abundant trinucleotide repeat in the sago palm nuclear genome, embodying 21.02% of the population. The aforesaid outcome matched with that of the

chloroplast genomes of *M. warburgii* and *M. sagu* (Barrett et al., 2016). As for oil and date palm, the most ascendant trinucleotide repeat is AAG/CTT (40%) and AGG/CCT (26.8%) respectively (Zhao et al., 2013; Kalyana et al., 2018). Intriguingly, the pentanucleotide and tetranucleotide repeats with the greatest frequency in the sago palm nuclear genome are extremely similar, namely AAAAT/ATTTT and AAAT/ATTT, they exemplify 11.98% and 16.86% of their respective populations. Circumstantially, the most copious tetranucleotide repeat in both chloroplast genomes of natangura and sago palm echoed that of the sago palm nuclear genome. Nonetheless, the pentanucleotides that are prevalent in both the aforesaid chloroplast genomes are AACTT/AAGTT and AAAGT/ACTTT instead (Barrett et al., 2016). The pentanucleotide and tetranucleotide data for both date and oil palm are not tapped into as yet. These SSR information generated from this study delivers an exceptional groundwork for the establishment of sago palm genome and molecular indicators henceforth.

4.4.5 Starch-related Homology Analysis

The starch-linked protein BLAST was conducted on the sago palm genome scaffolds to seek for the implications of these proteins in sago palm starch synthesis. A total of 238 sago palm genome scaffolds were uncovered to have interconnectedness with the key starch proteins observed in rice (*O. sativa*). Around 20% of these sago palm genome scaffolds code for the branching enzyme precursors. This enzyme is crucial in the orchestration of amylopectin biosynthesis and starch granule fabrication, contributing towards the physical traits and conformation of starch granules (Nakamura et al., 2010; Tetlow & Emes, 2014; Sawada et al., 2018). Approximately 10% of the sago palm genome scaffolds contain starch branching enzyme rbe4 gene. In rice, the expression of the rbe4

enzyme is the quickest in contrast to other starch synthesis enzymes (Mizuno et al., 2001; Ohdan et al., 2005). In wheat, a rise in the endosperm amylose substance was observed to be coupled with the plunge in the expression of *rbe4* (Regina et al., 2006).

In sago palm, the starch harvest from palms cultivated in mineral soil is much greater than that gathered from trees flourishing in acid sulfate soil (Nozaki et al., 2004; Rawung & Indrasti, 2021). The difference is more apparent at bottom parts of the trunk with regards to the starch yield content in a single sago palm trunk. It was also documented that sago palm thriving on mineral soil yielded tinier sized storage starch granules (Nozaki et al., 2004). The starch intensity and amount were regulated by granule-bound starch synthase, isoamylase as well as starch branching enzymes (Wang et al., 2018). The sago palm starch branching enzymes has a smaller action intensity in distinction to that of rice and maize. When being exposed to unpleasant conditions, sago palm inclines to employ storage elements to surmount the adverse impacts of regular development hindrances (Nozaki et al., 2004; Chua et al., 2021). Over this timeframe, the action of starch branching enzymes is raised remarkably (Sumardiono et al., 2021). The substantial roles of these starch-associated enzymes in sago palm in defining starch yield is principally the focus of sago palm genetic research, and this is one spotlight not to be overlooked on the voyage to crop enhancement through genetic engineering tools.

4.4.6 Identification of rRNA Cistron Components and Phylogenetic Analysis

The sago palm 28S and 18S rRNA sequences were excavated from the entire genome sequences and successively blasted against the NCBI GenBank public database. The sago palm 18S rRNA gene is sized at 1810 bp, which this size does not vary far from that of soybean, rice and maize (1807 bp, 1812 bp and 1809 bp correspondingly) (Takaiwa

et al., 1984; Messing et al., 1985; Dams et al., 1987). The sago palm 28S rRNA gene (4030 bp) though, is much bigger than that of the rice (3376 bp) (Takaiwa et al., 1985a). The sago palm 5.8S rRNA gene length revealed in the current study (160 bp) is only 3 bp smaller than that of its congeneric family member (*Metroxylon salomonense*), rice and broad bean (Tanaka et al., 1980; Takaiwa et al., 1985b). The leading nucleotide hits for both sago palm 28S and 18S rRNA are that of the *C. nucifera* (coconut palm). Incidentally, the similarity percentage and e-value were 98% and 0.0 correspondingly for both sago palm rRNA fragments. The probable justification for this result is that there is a shortage with regards to 28S and 18S sequence quantity deposited in the GenBank database, particularly within the palm family. The coconut palm and sago palm originate from varying subfamily and tribe but both of them are categorised under Arecaceae family, indicating the superior sequence preservation of 28S and 18S rRNA within the palm family.

The sago palm ITS1 (264 bp) is 10 bp smaller than that of the Solomon ivory nut palm (*M. salomonense*) (274 bp), 38 bp smaller than that of the oil palm (302 bp) and 46 bp smaller than that of the coconut palm (310 bp) (Baker et al., 2000; Eiserhardt et al., 2011). On the other hand, the sago palm ITS2 (249 bp) is 18 bp and 15 bp smaller than that of coconut and oil palm (267 bp and 264 bp) but 3 bp lengthier than that of *M. salomonense* (246 bp) (Baker et al., 2000; Eiserhardt et al., 2011). The ITS2 region was chosen for downstream phylogenetic tree building owing to its benefit of having both variable sequences and low intra-specific varied preserved secondary conformation (Han et al., 2013). The maximum likelihood phylogenetic tree in the current study illustrated a lush assemblage of 13 tribes from the palm family. Majority of them are categorised under the Arecoideae tribe and *Pinanga* is the most copious genus. An analogous phylogenetic tree has been built by Baker et al. (2000) over two decades ago employing the entire ITS region

by means of maximum parsimony paradigm. Intriguingly, the tree charted by Baker et al. (2000) is very akin to that assembled in the current study generally with several exclusions. The addition of extra 77 new palm species from five new tribes sequenced within the twenty years' timeframe not unearthed by Baker et al. (2000) but present in this study delivers new perceptions and substantiates the preceding assignments by Baker et al. (2000). One noteworthy resemblance worth emphasizing among both trees is that the Metroxylinae tribe has two closely spaced clusters housing members like *M. sagu* and *M. salomonense* as well as *K. jala*, *K. rogersii* and *K. cheb* correspondingly (*K. rogersii* is not present in Baker et al. (2000)). In short, this chapter has contributed towards the future sago genome assembly when sago palm Nanopore long sequencing reads are available in future to complement the Illumina sequencing reads generated in this study in a hybrid assembly that are indispensable in making the sago palm genome sequences more complete.

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

The preliminary genomic landscape of sago palm has been moulded with the achievement of all objectives set in this research project. The complete chloroplast genome of sago palm sequenced and characterised in this study is 157,300 bp long. Besides, the organellar genome copy number enumeration analysis had unearthed that the mature leaf has the most copious plastome and mitogenome copies in contrast to other organs tested. The *Plawei Mani* growth stage is also documented to be the most active developmental stage of sago palm with the highest plastome and mitogenome copy numbers. The trunking phenotype was proven to have higher metabolic requirements in terms of organellar genome copy number as compared to the spiny and non-trunking phenotypes. No significant difference in organellar genome copy number was detected across sago palms harvested from all Sarawak localities examined in this study. The sago palm nuclear genome was recorded at 616,599,600 bp in the current study and the identified microsatellites and gene homologs are valuable resources for future taxonomical and functional characterization research of the sago palm. The first research chapter yielded the complete chloroplast genome of the sago palm that is beneficial towards future taxonomical and metabolic pathway studies. The second research chapter documented on the organellar genome copy changes under various circumstances and these data is indispensable in terms of future gene expression studies. The final research chapter pathed the preliminary path in the future complete sago palm genome assembly when the sago palm Nanopore sequencing reads are available.

5.2 Recommendations

There were several limitations present in this study that require future attention and improvements for the advancement and completion of the whole sago palm genomic landscape. First, it is recommended that a larger sampling size be utilised in future investigation to further validate the outcomes in the present study. Furthermore, it is also proposed to include more sampling sites from other localities in Sarawak and even from other states in Peninsular Malaysia to further improve the comprehensiveness of the conclusion drawn in this study that the localities are not correlated with the organellar genome copy number variation. Lastly, a global sago palm chloroplast and nuclear genome comparison can be conducted in future for haplotype and genome-wide association studies.

REFERENCES

- Abbas, B. (2018). Sago palm genetic resource diversity in Indonesia. In H. Ehara, Y. Toyoda & D. Johnson (Eds.), *Sago palm* (pp. 61-74). Singapore: Springer.
- Abbas, B., Dailami, M., Santoso, B., & Munarti. (2017). Genetic variation of sago palm (*Metroxylon sagu* Rottb.) progenies with natural pollination by using RAPD markers. *Natural Science*, 9(4), 104-109.
- Abbas, B., Renwarin, Y., Bintoro, M. H., Sudarsono, Surahman, M., & Ehara, H. (2010). Genetic diversity of sago palm in Indonesia based on chloroplast DNA (cpDNA) markers. *Biodiversitas*, 11(3), 112-117.
- Abbas, B., Tjolli, I., & Munarti. (2020). Genetic diversity of sago palm (*Metroxylon sagu*) accessions based on plastid cpDNA *matK* gene as DNA barcoding. *Biodiversitas*, 21(1), 219-225.
- Abdullah, Mehmood, F., Shahzadi, I., Waseem, S., Mirza, B., Ahmed, I., & Waheed, M. T. (2019). Chloroplast genome of *Hibiscus rosa-sinensis* (Malvaceae): Comparative analyses and identification of mutational hotspots. *Genomics*, S0888-7543(19), 30062X.
- Abramoff, M. D., Magalhaes, P. J., & Ram, S. J. (2004). Image processing with ImageJ. *Biophotonics International*, 11(7), 36-42.
- Agus, F., & Subiksa, I. G. M. (2008). *Lahan gambut: Potensi untuk pertanian dan aspek lingkungan*. [Peatlands: Potential for agriculture and environment.] Bogor, Indonesia: Balai Penelitian Tanah dan World Agroforestry Centre (ICRAF).

- Ahlert, D., Ruf, S., & Bock, R. (2003). Plastid protein synthesis is required for plant development in tobacco. *Proceedings of the National Academy of Sciences*, 100(1), 15730-15735.
- Ahmad, M. (2013). Sago's role as food stock in 21st century. *International Journal on Advanced Science Engineering Information Technology*, 3(4), 15-17.
- Ahmad, M. N. (2015). *Production of cellobiose from sago frond as antimicrobial agent* (Master's progress report), Universiti Malaysia Sarawak, Malaysia.
- Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., & Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, 12(1), R18.
- Alexeyev, M., Shokolenko, I., Wilson, G., & LeDoux, S. (2013). The maintenance of mitochondrial DNA integrity: Critical analysis and update. *Cold Spring Harbor Perspectives in Biology*, 5(1), a012641.
- Al-Dous, E. K., George, B., Al-Mahmoud, M. E., Al-Jaber, M. Y., Wang, H., Salameh, Y. M., Al-Azwani, E. K., Chaluvadi, S., Pontaroli, A. C., DeBarry, J., Arondel, V., Ohlrogge, J., Saie, I. J., Suliman-Elmeer, K. M., Bennetzen, J. L., Kruegger, R. R., & Malek, J. A. (2011). *De novo* genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nature Biotechnology*, 29(1), 521-527.
- Al-Harrasi, I., Al-Yahyai, R., & Yaish, M. W. (2018). Differential DNA methylation and transcription profiles in date palm roots exposed to salinity. *PLoS ONE*, 13(1), e0191492.
- Al-Harrasi, A., Khan, A. L., Asaf, S., & Al-Rawahi, A. (2019). *Boswellia sacra* plastid genome sequencing and comparative analysis. In A. Al-harrasi, A. L. Khan. S. Asaf, &

- A. Al-Rawahi (Eds.), *Biology of Genus Bosweillia* (pp103-121). Switzerland: Springer.
- Al-Mssallem, I. S., Hu, S., Zhang, X., Lin, Q., Liu, W., Tan, J., Yu, X., Liu, J., Pan, L., Zhang, T., Yin, Y., Xin, C., Wu, H., Zhang, G., Abdullah, M. M., B., Huang, D., Fang, Y., Alnakhlhi, Y. O., Jia, S., Yin, A., Alhuzimi, E. M., Alsaihati, B. A., Al-Owayyed, S. A., Li, F., Tala, Wang, J., Yun, Q., Alnassar, N. A., Wang, L., Yang, M., Al-Jelaify, R. F., Liu, K., Gao, S., Chen, K., Alkhaldi, S. R., Liu, G., Zhang, M., Guo, H., & Yu, J., (2013). Genome sequence of the date palm *Phoenix dactylifera* L. *Nature Communications*, 4(1), 2274.
- Alias, M. I. (2012). Analysis of starch from trunking and non-trunking sago palm (Doctoral dissertation), Universiti Malaysia Sarawak, Malaysia.
- Al Kharusi, L., Al Yahyai, R., & Yaish, M. W. (2019). Antioxidant response to salinity in salt-tolerant and salt-susceptible cultivars of date palm. *Agriculture*, 9(1), 8.
- Al-Nakeeb, K., Peterson, T. N., & Sicheritz-Pontén, T. (2017). Norgal: Extraction and de novo assembly of mitochondrial DNA from whole-genome sequencing data. *BMC Bioinformatics*, 18(1), 510.
- Alverson, A. J., Zhuo, S., Rice, D. W., Sloan, D. B., & Palmer, J. D. (2011). The mitochondrial genome of the legume *Vigna radiata* and the analysis of recombination across short mitochondrial repeats. *PLoS ONE*, 6(1), e16404.
- Amiryousefi, A., Hyvönen, J., & Pocai, P. (2018). IRscope: An online program to visualize the junction sites of chloroplast genomes. *Bioinformatics*, 34(17), 3030-3031.

- Apel, W., & Bock, R. (2009). Enhancement of carotenoid biosynthesis in transplastomic tomatoes by induced lycopene-to-provitamin A conversion. *Plant Physiology*, 12(1), 1274-1285.
- Apse, M. P., Sottosanto, J. B., & Blumwald, E. B. (2003). Vacuolar cation/H⁺ exchange, ion homeostasis, and leaf development are altered in a T-DNA insertional mutant of AtNHX1, the Arabidopsis vacuolar Na⁺/H⁺ antiporter. *The Plant Journal: for Cell and Molecular Biology*, 36(2), 229-239.
- Arrieta-Montiel, M., Lyznik, A., Wlozynska, M., Janska, H., Tohme, J., & Mackenzie, S. (2001). Tracing evolutionary and developmental implications of mitochondrial stoichiometric shifting in the common bean. *Genetics*, 158(1), 851-864.
- Arrieta-Montiel, M. P., Shedge, V., Davila, J., Christensen, A. C., & Mackenzie, S. A. (2009). Diversity of the *Arabidopsis* mitochondrial genome occurs via nuclear-controlled recombination activity. *Genetics*, 183(1), 1261-1268.
- Asaf, S., Wagas, M., Khan, A. L., Khan, M. A., Kang, S. M., Imran, Q. M., Shahzad, R., Bilal, S., Yun, B. W., & Lee, I. J. (2017). The complete chloroplast genome of wild rice (*Oryza minuta*) and its comparison to related species. *Frontiers in Plant Science*, 8(1), 304.
- Awg-Adeni, D. S., Abd-Aziz, S., & Bujang, K., & Hassan, M. A., (2010). Bioconversion of sago residue into value added products. *African Journal of Biotechnology*, 9(14), 2016-2021.
- Awg-Adeni, D. S., Ahmad, M. N., & Bujang, K. (2018). Maximising production of prebiotic sugar (cellobiose) from sago frond. *Malaysian Applied Biology*, 47(1), 89-95.

- Awg-Adeni, D. S., Bujang, K., Hassan, M. A., & Abd-Aziz, S. (2013). Recovery of glucose from residual starch of sago *hampas* for bioethanol production. *BioMed Research International*, 2013(1), 1-8.
- Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* (pp. 28-36), Menlo Park, CA: AAAI Press.
- Baker, W. J., Hedderson, T. A., & Dransfield, J. (2000). Molecular phylogenetics of subfamily Calamoideae (Palmae) based on nrDNA ITS and cpDNA rps16 intron sequence data. *Molecular Phylogenetics and Evolution*, 14(2), 195-217.
- Bakker, F. T., Culham, A., Gomez-Martinez, R., Carvalho, J., Compton, J., Dawtrey, R., & Gibby, M. (2000). Patterns of nucleotide substitution in angiosperm cpDNA trnL (UAA)-trnF(GAA) regions. *Molecular Biology and Evolution*, 17(8), 1146-1155.
- Banu, J. R., Kaliappan, S., & Beck, D. (2006). High rate anaerobic treatment of sago wastewater using HUASB with PUF as carrier. *International Journal of Environmental Science and Technology*, 3(1), 69-77.
- Barrett, C. F., Baker, W. J., Comer, J. R., Conran, J. G., Lahmeyer, S. C., Leebens-Mack, J. H., Li, J., Lim, G. S., Mayfield-Jones, D. R., Perez, L., Medina, J., Pires, J. C., Santos, C., Stevenson, D. W., Zomlefer, W. B., & Davis, J. I. (2016). Plastid genomes reveal support for deep phylogenetic relationships and extensive rate variation among palms and other commelinid monocots. *New Phytologist*, 209(2), 855-870.
- Beier, S., Thiel, T., Münch, T., Scholz, U., & Macher, M. (2017). MISA-web: A web server for microsatellite prediction. *Bioinformatics*, 33(1), 2583-2585.

- Bendich, A. J. (1987). Why do chloroplasts and mitochondria contain so many copies of their genome? *Bioessays*, 6(1), 279-282.
- Bengtsson-Palme, J., Veldre, V., Ryberg, M., Hartmann, M., Branco, S., Wang, Z., Godhe, A., De Wit, P., Sanchez-Garcia, M., Ebersberger, I., de Sousa, F., Amend, A., Jumpponen, A., Unterseher, M., Kristiansson, E., Abarenkov, K., Bertrand, Y. J. K., Sanli, K., Eriksson, K. M., Vik, U., Veldre, V., & Nilsson, R. H. (2013). ITSx: Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequence of fungi and other eukaryotes for use in environmental sequencing. *Methods in Ecology and Evolution*, 4(1), 914-919.
- Bhagyawant, S. S. (2016). RAPD-SCAR markers: An interface tool for authentication of traits. *Journal of Bioscience and Medicines*, 4(1), 1-9.
- Bintoro, H. M. H. (2008). *Bercocok tanam sagu*. [Sago plantation.] Bogor, Indonesia: IPB Press.
- Bintoro, H. M. H., Nurulhaq, M. I., Pratama, A. J., Ahmad, F., & Ayulia, L. (2018). Growing area of sago palm and its environment. In H. Ehara, Y. Toyoda, & D. Johnson (Eds.), *Sago palm: Multiple contributions to food security and sustainable livelihoods* (pp. 17-29). Singapore: Springer.
- Bintoro, H. M. H., Purwanto, Y. J., & Amarillis, S. (2010). *Sagu di lahan gambut*. [Sago in peatlands.] Bogor, Indonesia: IPB Press.
- Bock, R. (2000). Sense from nonsense: How the genetic information of chloroplasts is altered by RNA editing. *Biochimie*, 82(1), 549-557.
- Boesch, P., Weber-Lotfi, F., Ibrahim, N., Tarasenko, V., Cosset, A., Paulus, F., Lightowlers, R. N., & Dietrich, A. (2011). DNA repair in organelles: Pathways,

- organization, regulation, relevance in disease and aging. *Biochimica et Biophysica Acta*, 1813(1), 186-200.
- Boffey, S. A., & Leech, R. M. (1982). Chloroplast DNA levels and the control of chloroplast division in light-grown wheat leaves. *Plant Physiology*, 69(6), 1387-1391.
- Boonsermsuk, S., Hisajima, S., & Ishizuka, K. (1995). Isozyme variation in sago palm (*Metroxylon* spp.) in Thailand. *Japanese Journal of Tropical Agriculture*, 39(4), 229-235.
- Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1), 59-60.
- Bujang, K. B. (2015). Production, purification and health benefits of sago sugar. In E. Ehara, Y. Toyoda, & T. Mishima (Eds.), *Proceedings of the 12th International Sago Symposium: The Sago Supports Human Planet Welfare* (pp. 1-4). Tokyo, Japan: The Sago Palm Studies.
- Bujang, K. B. (2018). Production, purification, and health benefits of sago sugar. In H. Ehara, Y. Toyoda, & D. Johnson (Eds.), *Sago palm: Multiple contributions to food security and sustainable livelihoods* (pp. 299-307). Singapore: Springer.
- Bujang, K. B., & Jobli, S. (2002). Effects of glucose feed concentrations on continuous lactate production from sago starch. *International Symposium Tropical Resources and Green Chemistry Strategy* (pp. 358-363). Osaka, Japan: ICBiotech.
- Bujang, K. B., Apun, K., & Salleh, M. A. (1996). A study in the production and bioconversion of sago waste. *Sago-The Future Source of Food and Feed. Riau University Press, Indonesia*, 1(1), 195-201.

- Bujang, K. B., Awg-Adeni, D. S., & Jolhiri, P. (2000a). Effects of starch concentration and pH on enzymatic hydrolysis of sago starch. *Biotechnology for sustainable utilization of biological resources in the tropics* (pp. 32-35). Osaka, Japan: ICBiotech.
- Bujang, K. B., Bujang, S., & Awg-Adeni, D. S. (2004). *Effects of calcium carbonate in fermentation of L-lactic acid from hydrolysed sago starch*. Retrieved December 25, 2020, from <http://www.icb.osaka-u.ac.jp/AnnuRep/AnnuRep26/637-643.pdf>
- Bukhari, N. A., Soh, K. L., Bakar, N. A., & Ismail, M. (2017). Hydrolysis of residual starch from sago pith residue and its fermentation to bioethanol. *Sains Malaysiana*, 46(8), 1269-1278.
- Bustin, S. A., Benes, V., Garson, J. A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M. W., Shipley, G. L., Vandesompele, J., & Wittwer, C. T. (2009). The MIQE guidelines: Minimum information for publication of quantitative real-time PCR experiments. *Clinical Chemistry*, 55(1), 611-622.
- Campbell, B. C., Gilding, E. K., Mace, E. S., Tai, S., Tao, Y., Prentis, P.J., Thomelin, P., Jordan, D. R., & Godwin, I. D. (2016). Domestication and the storage starch biosynthesis pathway: Signatures of selection from a whole sorghum genome sequencing strategy. *Plant Biotechnology Journal*, 14(12), 2240-2253.
- Che Jusoh, M. R., Stannard, S. R., & Mündel, T. (2016). Sago supplementation for exercise performed in a thermally stressful environment: Rationale, efficacy and opportunity. *Temperature*, 3(3), 384-393.
- Chew, T. A., Abu Hassan, M. I., & Mohd Ghazali, M. (1999). The sago industry in Malaysia: Present status and future prospects. In: *Proc. 7th International Working Conference on stored-product protection volume 2* (pp. 1720–1728).

- Chen, H., Keseler, I. M., & Shimkets, L. J. (1990). Genome size of *Myxococcus xanthus* determined by pulsed field gel electrophoresis. *Journal of Bacteriology*, 172(1), 4206-4213.
- Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17), i884-i890.
- Chen, W. B., Hasegawa, D. K., Arumuganathan, K., Simmons, A. M., Wintermantel, W. M., Fei, Z. J., & Ling, K. S. (2015). Estimation of the whitefly *Bemisia tabaci* genome size based on kmer and flow cytometric analyses. *Insects*, 6, 704715.
- Chua, S. N. D., Kho, E. P., Lim, S. F., & Hussain, M. H. (2021). Sago palm (*Metroxylon sagu*) starch yield, influencing factors and estimation from morphological traits. *Advances in Materials and Processing Technologies*, DOI: 10.1080/2374068X.2021.1878702.
- Chumley, T. W., Palmer, J. D., Mower, J. P., Fourcade, H. M., Calie, P. J., Boore, J. L., & Jansen, R. K. (2006). The complete chloroplast genome sequence of *Pelargonium x hortorum*: Chloroplast genome of land plants. *Molecular Biology and Evolution*, 23(11), 2175-2190.
- Chung, H. J., Jung, J. D., Park, H. W., Kim, J. H., Cha, H. W., Min, S. R., Jeong, W. J., & Liu, J. R. (2006). The complete chloroplast genome sequences of *Solanum tuberosum* and comparative analysis with Solanaceae species identified the presence of a 241-bp deletion in cultivated potato chloroplast DNA sequence. *Plant Cell Reports*, 25(1), 1369-1379.
- Clegg, M. T., Gaut, B. S., Learn, Jr., G. H., & Morton, B. R. (1994). Rates and patterns of chloroplast DNA evolution. *The Proceedings of the National Academy of Sciences*, 91(1), 6795-6801.

- Daniell, H., Lin, C. S., Yu, M., & Chang, W. J. (2016). Chloroplast genomes: diversity, evolution, and application in genetic engineering. *Genome Biology*, 17(1), 134.
- Darwesh, R. S. S. (2013). Improving growth of date palm plantlets grown under salt stress with yeast and amino acids applications. *Annals of Agricultural Science*, 58(2), 247-256.
- Davila, J. I., Arrieta-Montiel, M. P., Wamboldt, Y., Cao, J., Hagmann, J., Shedge, V., Xu, Y. Z., Weigel, D., & Mackenzie, S. A. (2011). Double-strand break repair processes drive evolution of the mitochondrial genome in *Arabidopsis*. *BMC Biology*, 9(1), 64.
- Davoodi-Semiromi, A., Schreiber, M., Nalapali, S., Verma, D., Singh, N. D., Banks, R. K., Chakrabarti, D., & Daniell, H. (2010). Chloroplast-derived vaccine antigens confer dual immunity against cholera and malaria by oral or injectable delivery. *Plant Biotechnology Journal*, 8(2), 223-242.
- de Castro, E., Sigrist, C.J.A., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A. & Hulo, N. (2006). ScanProsite: Detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Research*, 34(1), W362- W365.
- De Vita, R., Cavallo, D., Eleuteri, P., Odierna, G., & Balletto, E. (1997). Flow cytometric approach to study genome size variation in eurasiatic green toads of the *Bufo viridis* complex. *European Journal of Histochemistry*, 41(1), 175-176.
- Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). NOVOPlasty: *de novo* assembly of organellar genomes from whole genome data. *Nucleic Acids Research*, 45(4), e18.
- Djoefrie, H. M. H. B. (1999). *Pemberdayaan tanaman sagu sebagai penghasil bahan pangan alternative dan bahan baku agroindustry yang potensial dalam rangka*

ketahanan pangan nasional. [Empowerment of sago plantation as alternative food and agroindustrial raw material in the national food security program.] Bogor, Indonesia: Orasi Ilmiah Guru Besar Tetap Ilmu Tanaman Perkebunan Fakultas Pertanian Institut Pertanian Bogor.

Donaldson, J. (2003). Cycads of the world: An overview. In J. Donaldson (Ed.), *Cycads: Status survey and conservation action plan* (pp. 3-7). Cambridge, United Kingdom: International Union for Conservation of Nature.

Dong, S., Zhao, C., Chen, F., Liu, Y., Zhang, S., Wu, H., Zhang, L., & Liu, Y. (2018). The complete mitochondrial genome of the early flowering plant *Nymphaea colorata* is highly repetitive with low recombination. *BMC Genomics*, 19(614), doi:10.1186/s12864-018-4991-4

Dong, W., Liu, H., Xu, C., Zuo, Y., Chen, Z., & Zhou, S. (2014). A chloroplast genome strategy for designing taxon specific DNA mini-barcodes: A case study on ginsengs. *BMC Genetics*, 15(1), 138.

Dong, W., Liu, J., Yu, J., Wang, L., & Zhou, S. (2012). Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *PLoS ONE*, 7(4), e35071.

Dong, W., Xu, C., Cheng, T., Lin, K., & Zhou, S. (2013). Sequencing angiosperm plastid genomes made easy: A complete set of universal primers and a case study on the phylogeny of saxifragales. *Genome Biology and Evolution*, 5(5), 989-997.

Dransfield, J. (1976). Dry land sago palms. In Tan K. (Ed.), *Sago-76, The Proceedings of the 1st Sago Symposium*, Sarawak, Malaysia.

- Dufay, M., Touzat, P., Maurice, S., & Cuguen, J. (2007). Modelling the maintenance of male-fertile cytoplasm in a gynodioecious population. *Heredity*, 99(1), 349-356.
- Edward, A. S. (2013). *Differential expression gene profiling studies on trunking and non-trunking sago palm (Metroxylon sagu Rottb.)* (Master's thesis), Universiti Malaysia Sarawak, Malaysia.
- Ehara, H. (2009). Potency of sago palm as carbohydrate resource for strengthening food security program. *Indonesian Journal of Agronomy*, 37(3), 209-219.
- Ehara, H. (2018). Genetic variation and agronomic features of *Metroxylon* palms in Asia and Pacific. In H. Ehara, Y. Toyoda, D. V. Johnson (Eds.), *Sago palm: Multiple contributions to food security and sustainable livelihoods*. Springer: Japan.
- Ehara, H., Matsui, M., & Naito, H. (2003). Absorption and translocation of Na^+ in sago palm under NaCl treatments. *Sago Palm*, 11(1), 35-36.
- Ehara, H., Matsui, M., & Naito, H. (2006). Avoidance mechanism of salt stress in sago palm (*Metroxylon sagu* Rottb.). *Japanese Society for Tropical Agriculture*, 50(1), 36-41.
- Ehara, H., Prathumyot, W., & Naito, H. (2011). Salt resistance of *Metroxylon sagu*, starch-producing palm. *Proceedings of the 7th ACSA Conference* (pp. 5-10). Bogor, Indonesia: Bogor Agricultural University.
- Ehara, H., Shibuya, H., Prathumyot, W., Naito, H., & Miyake, H. (2008). Absorption and distribution of Na^+ , Cl^- and some other ions and physiological characteristics of sago palm under salt stress. *Tropical Agriculture and Development*, 52(1), 7-16.
- Eiserhardt, W. L., Pintaud, J. C., Asmussen-Lange, C., Hahn, W. J., Bernal, R., Balslev, H., & Borchsenius, F. (2011). Phylogeny and divergence times of Bactridinae

- (Arecaceae, Palmae) based on plastid and nuclear DNA sequences. *Taxon*, 60(2), 485-498.
- Enguito, R., & Novero, A. (2018). Genetic variation of sago palm (*Metroxylon sagu* Rottb.) samples from Mindanao, Philippines using internal transcribed spacer analysis. *Malaysian Applied Biology*, 47(3), 39-48.
- Evans, D. L., Hlongwane, T. T., Joshi, S. V., & Pachón, D. M. R. (2019). The sugarcane mitochondrial genome: Assembly, phylogenetics and transcriptomics. *PeerJ*, 7(1), e7558.
- El Rabey, H. A., Al-Malki, A. L., & Abulnaja, K. O. (2016). Proteome analysis of date palm (*Phoenix dactylifera* L.) under severe drought and salt stress. *International Journal of Genomics*, 2016(1), 1-8.
- Faiyue, B., Al-Azzawi, M. J., & Flowers, T. J. (2012). A new screening technique for salinity resistance in rice (*Oryza sativa* L.) seedlings using bypass flow. *Plant Cell Environment*, 35(1), 1099-1108.
- Flach, M. (1977). Yield potential of the sago palm and its realization. In K. Tan (Ed.), *Sago-'76: Papers of the 1st International Sago Symposium* (pp. 157-177). Kuala Lumpur, Malaysia: University of Malaya.
- Flach, M., Braber, K. D., Fredrix, M. J. J., Monster, E. M., & van Hasselt, G. A. M. (1986). Temperature and relative humidity requirements of young sago palm seedlings. In N. Yamada, & K. Kainuma (Eds.), *Proceedings of the Third International Sago Symposium* (pp. 139-143). Tokyo, Japan: Sago Palm Research Fund.
- Flach, M., & Schuling, D. L. (1991). Growth and yield of sago palms in relation to their nutritional needs. In T. T. Ng, Y. L. Tie, & H. S. Kueh (Eds.), *Proceedings of the*

- fourth international sago symposium* (pp. 103-110). Sarawak, Malaysia: Lee Ming Press.
- Fleischmann, A., Michael, T. P., Rivadavia, F., Sousa, A., Wang, W. Q., Temsch, E. M., Greilhuber, J., Muller, K. F., & Heubl, G. (2014). Evolution of genome size and chromosome number in the carnivorous plant genus *Genlisea* (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. *Annals of Botany*, 114(8), 1651-1663.
- Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., & Dubchak, I. (2004). VISTA: Computational tools for comparative genomics. *Nucleic Acids Research*, 32(1), W273-W279.
- Frenkel, N., & Roizman, B. (1971). Herpes simplex virus: Genome size and redundancy studied by renaturation kinetics. *Journal of Virology*, 8(1), 591-593.
- Gao, J., & Scott, J. G. (2006). Use of quantitative real-time polymerase chain reaction to estimate the size of the house-fly *Musca domestica* genome. *Insect Molecular Biology*, 15(6), 835-837.
- Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D. & Bairoch, A. (2005). Protein identification and analysis tools on the ExPASy server. In J. M. Walker (Ed.), *The Proteomics Protocols Handbook*, 1st edition (pp. 571-607). Totowa, NJ: Humana Press.
- Gatehouse, J. A. (1997). Regulation of gene expression in plants. In P. M. Dey & J. B. Harborne (Eds.), *Plant Biochemistry* (pp. 353-385). Cambridge, MA: Academic Press.
- Gene Codes. (2019). Sequencher (version 4.10.1). Retrieved from <http://www.genecodes.com>

- Ghosh, A. K., Rahaman, A. A., & Singh, R. (2010). Combination of sago and soy-protein supplementation during endurance cycling exercise and subsequent high-intensity endurance capacity. *International Journal of Sport Nutrition and Exercise Metabolism*, 20(1), 216-223.
- Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R. L., & Dunn, M. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). *Science*, 296(1), 92-100.
- Gong, H., Randall, D., & Flowers, T. (2006). Silicon deposition in the root reduces sodium uptake in rice (*Oryza sativa* L.) seedlings by reducing bypass flow. *Plant Cell Environment*, 29(1), 1970-1979.
- Gray, M. W. (2004). The evolutionary origins of plant organelles. In H. Daniell, & C. Chase (Eds.), *Molecular biology and biotechnology of plant organelles chloroplast and mitochondria* (pp. 15-36). Dordrecht: Springer.
- Greenhill, A. R. (2006). *Food safety and security of sago starch in rural Papua New Guinea* (Doctoral dissertation). James Cook University, Townsville, Queensland, Australia.
- Gregory, T. R., Nathwani, P., Bonnett, T. R., & Huber, D. P. W. (2013). Sizing up arthropod genomes: An evaluation of the impact of environmental variation on genome size estimates by flow cytometry and the use of qPCR as a method of estimation. *Genome*, 56(9), 505-510.
- Gu, Y., Han, S., Chen, L., Mu, J., Duan, L., Li, Y., Yan, Y., & Li, X. (2021). Expression and regulation of genes involved in the reserve starch biosynthesis pathway in hexaploid wheat (*Triticum aestivum* L.). *The Crop Journal*, 9(2), 440-455.

- Gualberto, J. M., Mileschina, D., Wallet, C., Niazi, A. K., Weber-Lotfi, F., & Dietrich, A. (2014). The plant mitochondrial genome: Dynamics and maintenance. *Biochimie*, 100(1), 107-120.
- Gualberto, J. M., & Newton, K. J. (2017). Plant mitochondrial genomes: dynamics and mechanisms of mutation. *Annual Review of Plant Biology*, 68(13), 1-28.
- Guisinger, M. M., Chumley, T. W., Kuehl, J. V., Boore, J. L., & Jansen, R. K. (2010). Implications of the plastid genome sequence of *Typha* (Typhaceae, Poales) for understanding genome evolution in Poaceae. *Journal of Molecular Evolution*, 70(1), 149-166.
- Gunawan, S., Apamarta, H. W., Zarkasie, I. M., & Prihandini, W. W. (2018). Effect of initial bacterial cells number and fermentation time on increasing nutritive value of sago flour. *Malayian Journal of Fundamental and Applied Sciences*, 14(2), 246-250.
- Gurevich, A., Savaliev, V., Vyahhi, N., & Tesler, G. (2013). QUASt: Quality assessment tool for genome assemblies. *Bioinformatics*, 29(8), 1072-1075.
- Han, J., Zhu, Y., Chen, X., Liao, B., Yao, H., Song, J., Chen, S., & Meng, F. (2013). The short ITS2 sequence serves as an efficient taxonomic sequence tag in comparison with the full-length ITS. *BioMed Research International*, 2013(1), 741476.
- Hamanishi, T., Hatta, T., Jong, F.S., Takahashi, S., & Kainuma, K. (1999). Physicochemical properties of starches obtained from various parts of sago palm trunks at different growth stages. *Journal of Applied Glycoscience*, 46(1), 39-48.
- Hamanishi, T., Hatta, T., Jong, F. S., Kainuma, K., & Takahashi, S. (2000). The relative crystallinity, structure and gelatinization properties of sago starches at different growth stages. *Journal of Applied Glycoscience*, 47(3), 335-41.

- Hamilton, M. B., Braverman, J. M., & Soria-Hernanz, D. F. (2003). Patterns and relative rates of nucleotide and insertion/deletion evolution of six chloroplast intergenic regions in new world species of the Lecythidaceae. *Molecular Biology and Evolution*, 20(10), 10-21.
- Hammer, Ø., Harper, D. A. T., & Ryan, P. D. (2001). PAST: Paleontological statistics software package for education and data analysis. *Palaeontologia Eelectronica*, 4(1), 1-9.
- Haryanto, B., & Pangloli, P. (1992). *Potensi dan pemanfaatan sagu*. [The potential and benefits of sago.] Jogjakarta, Indonesia: Penerbit Kanisius Jogjakarta.
- Heslop-Harrison, J. S. (2003). Crop improvement plant genomes. In B. Thomas (Ed.), *Encyclopedia of Applied Plant Sciences* (pp. 130-133). Amsterdam, Netherlands: Elsevier.
- Hidayat, S., Matsuoka, M., Baja, S., & Rampisela, D. A. (2018). Object-based image analysis for sago palm classification: The most important features from high-resolution satellite imagery. *Remote Sensing*, 10(8), 1319.
- Hipolito, C. N., Crabbe, E., Kobayashi, G., Sonomoto, K., & Ishizaki, A. (2000). pH-dependent continuous lactic acid fermentation by *Lactococcus lactis* IO-1 using hydrolysed sago starch. *Journal of the Faculty of Agriculture, Kyushu University*, 44(3-4), 367-375.
- Hipolito, C. N., Zarrabal, O. C., Kamaldin, R. M., Yee, L. T., Lihan, S., Bujang, K., & Nitta, Y. (2012). Lactic acid production by *Enterococcus faecium* in liquified sago starch. *AMB Express*, 2(1), 53.

- Hirao, K., Kondo, T., Kainuma, K., & Takahashi, S. (2018). Starch properties and uses as food for human health and welfare. In H. Ehara, Y. Toyoda, & D. Johnson (Eds.), *Sago palm* (pp. 285-298). Springer: Singapore.
- Hirao, K., & Takanashi, S. (1996). Studies on cooking of pearl-type starch part 3: Cooking method of sago pearls. *Sago Palm*, 4(1), 14-20.
- Hirao, K., Tanaka, H., & Konoo, S. (2008). The present condition of sago starch using method and processed foods in Riau, Indonesia. *Proceedings of the 17th conference of the society of sago palm studies* (pp. 75-78). Tokyo, Japan: Japanese Society Sago Palm Studies.
- Hirotsu, S., Okazaki, M., Toyota, K., Yano, Y., Seki, K., Tsujimoto, H., Ham, Y. S., & Power, P. (2002). Determination of osmoregulating compounds in sago palm. *Abstracts of the 11th conference of the Japanese Society Sago Palm Studies* (pp. 11-16). Tokyo, Japan: Japanese Society Sago Palm Studies.
- Holm, L. & Laakso, L. M. (2016). Dali server update. *Nucleic Acids Research*, 44 (W1): W351-W355.
- Huang, Y. Y., Matzke, A. J., & Matzke, M. (2013). Complete sequence and comparative analysis of the chloroplast genome of coconut palm (*Cocos nucifera*). *PLoS ONE*, 8(8), e74736.
- Husaini, A. A. S. A., Hussain, M. H. M., & Roslan, H. A. (2016). Sago palm (*Metroxylon sagu* Rottb.): A potential source of starch for food security. In J. Mohd-Azlan & I. Das (Eds.), *Biodiversity of tropical peat swamp forests of Sarawak* (pp. 81-114). Sarawak, Malaysia: UNIMAS Publisher.

- Hussain, H., Kamal, M. M., Atit, A. S. E., Kamarol, S. I. L., Yan, W. J., Ngaini, Z., Al-Obaidi, J. R., Mohd-Yusuf, Y., & Mohd-Yusuf, Y. (2012a). Proteomics of trunking and non-trunking sago palm (*Metroxylon sagu* Rottb.). In N. Busri, Z. C. Abdullah, C. N. Hipolito, H. A. Roslan, E. U. H., Sim, D. S. Awg-Adeni, A. A. S. A. Husaini, & N. Humrawali (Eds.), *Proceedings of 2nd ASEAN Sago Symposium 2012: Advances in Sago Research and Development* (p. 109). Sarawak, Malaysia: Universiti Malaysia Sarawak.
- Hussain, H., Kamal, M.M., Al-Obaidi, J.R., Hamdin, N.E., Ngaini, Z., & Mohd-Yusuf, Y. (2020). Proteomics of sago palm towards identifying contributory proteins in stress-tolerant cultivar. *The Protein Journal*, 39(1), 62–72.
- Hussain, H., Yan, W. J., & Ngaini, Z. (2012b). Spectroscopy profiling of trunking sago palm (*Metroxylon sagu* Rottb.) using nuclear magnetic resonance (NMR). In H. Sani, F. B. Ahmad, M. E. Wasli, P. T. Lim, K. S. Lee & Y. B. Esa (Eds.), *4th Regional Conference on Natural Resources in the Tropics*. Sarawak, Malaysia: Universiti Malaysia Sarawak.
- Initiative, A. G. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408(1), 796815.
- Inukai, T. (2017). Differential regulation of starch-synthetic gene expression in endosperm between *indica* and *japonica* rice cultivars. *Rice (N. Y.)*, 10(1), 7.
- Ishiaku, U. S., Pang, K. W., Lee, W. S., & Mohamad, I. Z. A. (2002). Mechanical properties and enzymic degradation of thermoplastic and granular sago starch filled polycaprolactone. *European Polymer Journal*, 38(2), 393–401.

- Iskow, R. C., Gokcumen, O., & Lee, C. (2012). Exploring the role of copy number variants in human adaptation. *Trends in Genetics*, 28(6), 245-257.
- Jamaludin, K., Paridah, T. M., Hamami, S., & Yin, W. T. (1995). Soda anthraquinone pulping of sago palm (*Metroxylon sagu* Rottboel) fronds. *Acta Horticulture*, 389 (fifth international sago symposium, 1994), 237–247.
- Jamel, B., Hussain, M. H., Salleh, M. A., & Busri, N. (2011). Isolation and characterization of the GA 20-oxidase cDNA from sago palm. *Asia-Pacific Journal of Molecular Biology and Biotechnology*, 19(2), 83-93.
- Jangu, U., & Bujang, K. B. (2009). Maximizing sugar production from enzymatic hydrolysis of sago fibre for ethanol fermentation. In *Proceedings of the 1st ASEAN Sago Symposium* (pp. 30-33). Sarawak, Malaysia: Universiti Malaysia Sarawak.
- Jin, S., & Daniell, H. (2015). The engineered chloroplast genome just got smarter. *Trends in Plant Science*, 20(1), 622-640.
- Johnravinder, D., Murugesen, K., Wong, J. W. C., & Elangovan, N. (2017). Waste-to biofuel: Production of biobutanol from sago waste residues. *Environmental Technology*, 38(13-14), 1725-1734.
- Johnson, D. (1977). *Distribution of sago making in the old world*. In K. Tan (Ed.), Sago-76: Papers of the First International Sago Symposium. Kuala Lumpur, Malaysia: Kemajuan Kanji Sdn. Bhd., 65-75.
- Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, 30(9), 1236-1240.

- Kadirvelu, K., Karthika, C., Vennilamani, N., & Patabhi, S. (2005). Activated carbon from industrial solid waste as an adsorbent for the removal of Rhodamine B from aqueous solution: Kinetic and equilibrium studies. *Chemosphere*, 60(8), 1009–1017.
- Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., Kohara, Y., Fujiyama, A., Hayashi, T., & Itoh, T. (2014). Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*, 24(1), 1384-1395.
- Kaku, S. (1996). Salt stress. In Y. Matsuda & M. Kikuyama (Eds.), *Plant physiology* (pp. 104-105). Tokyo, Japan: The Society for the Promotion of the University of the Air.
- Kalyana, B. B., Mary Rani, K. L., Sahu, S., Mathur, R. K., Naveen, K. P., Ravichandran, G., Anitha, P., & Bhagya, H. P. (2019). Development and validation of whole genome-wide and genic microsatellite markers in oil palm (*Elaeis guineensis* Jacq.): First microsatellite database (OpSatdb). *Scientific Reports*, 9(1), 1899.
- Kamal, M. M. (2014). Identification of differentially expressed protein in trunking and non-trunking sago palm leaves. In Universiti Malaysia Sarawak (Ed.), *Sago*. Sarawak, Malaysia: Universiti Malaysia Sarawak.
- Kanehsia, M., Sato, Y., & Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and mitogenome sequences. *Journal of Molecular Biology*, 428(1), 726-731.
- Karim, A. A., Tie, A. P. L., Manan, D. M. A., & Zaidul, I. S. M. (2008). Starch from the sago (Metroxylon sago) palm tree: Properties, prospects, and challenges as a new industrial source for food and other uses. *Comprehensive Reviews in Food Science and Food Safety*, 7(3), 215-228.

- Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, 10(6), 845- 858.
- Kiew, R. (1977). *Taxonomy, ecology and biology of sago palms in Malaya and Sarawak*. In K. Tan (Ed.), *Sago-76: Papers of the First International Sago Symposium*. Kuala Lumpur, Malaysia: Kemajuan Kanji Sdn. Bhd., 147–154.
- Kimura, N. (1979). Pests of sago palm and their control. *Japanese Journal of Tropical Agriculture*, 23(3), 148-156.
- Kjær, A., Barfod, A. S., Amussen, C. B., & Severg, O. (2004). Investigation of genetic and morphological variation in the sago palm (*Metroxylon sagu*; Arecaceae) in Papua New Guinea. *Annals of Botany*, 94(1), 109-117.
- Kokot, M., Długosz, M., & Deorowicz, S. (2017). KMC 3: Counting and manipulating k-mer statistics. *Bioinformatics*, 33(17), 2759-2761.
- Kong, S. L., Abdullah, S. N. A., Ho, C. L., Musa, M. H., & Yeap, W. C. (2021). Comparative transcriptome analysis reveals novel insights into transcriptional responses to phosphorus starvation in oil palm (*Elaeis guineensis*) root. *BMC Genomic Data*, 22(1), 6.
- Kueh, H. S. (1977). *Sago agronomy in Sarawak*. In K. Tan (Ed.), *Sago-76: Papers of the First International Sago Symposium*. Kuala Lumpur, Malaysia: Kemajuan Kanji Sdn. Bhd., 181–185.
- Kueh, H. S., Elone, R., Tie, Y. L., Ung, C. M., & Osman, J. (1987). *Mukah sago plantation feasibility study: Final report*. Sarawak, Malaysia: Lembaga Pembangunan dan Lindungan Tanah.

- Kumar, R. A., Oldenburg, D. J., & Bendich, A. J. (2014). Changes in DNA damage, molecular integrity, and copy number for plastid DNA and mitochondrial DNA during maize development. *Journal of Experimental Botany*, 65(22), 6425-6439.
- Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, 35(6), 1547-1549.
- Kumaran, S., Sastry, C. A., & Vickineswary, S. (1997). Laccase, cellulose and xylanase activities during growth of *Pleurotus sajor-caju* on sago hampas. *World Journal of Microbiology and Biotechnology*, 13(1), 43–49.
- Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., & Giegerich, R. (2001). REPuter: The manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research*, 29(1), 4633-4642.
- Larkin, R. M. (2019). RNA editing implicated in chloroplast-to-nucleus communication. *The Proceedings of the National Academy of Sciences*, 116(20), 9701-9703.
- Laskowski, R.A., Watson, J.D. & Thornton, J.M. (2005). ProFunc: A server for predicting protein function from 3D structure. *Nucleic Acids Research*, 33(suppl2), W89–W93.
- Lee, G. M., Kim, C. H., Lee, K. J., Yusof, Z. A. M., Han, M. H., & Rhee, S. K. (1986). Simultaneous saccharification and ethanol fermentation of sago starch using immobilized *Zymomonas mobilis*. *Journal of Fermentation Technology*, 64(4), 293-297.
- Lenz, H., Hein, A., & Knoop, V. (2018). Plant organelle RNA editing and its specificity factors: Enhancements of analyses and new database features in PREPACT 3.0. *BMC Bioinformatics*, 19(1), 255.

- Li, D. M., Zhao, C. Y., & Liu, X. F. (2019). Complete chloroplast genome sequences of *Kaempferia galanga* and *Kaempferia elegans*: Molecular structures and comparative analysis. *Molecules*, 24(3), E474.
- Li, G., Song, L., Jin, C., Li, M., Gong, S., & Wang, Y. (2019). Genome survey and SSR analysis of *Apocynum venetum*. *Bioscience Reports*, 39(1), 1-11.
- Li, L. B., Wu, J. Y., Hu, T., Yang, X. W., & Peng, Z. H. (2008). Estimation of genome size of Moso Bamboo (*Phyllostachys edulis*). *Chinese Bulletin of Botany*, 25(1), 574-578.
- Li, W., Ruf, S., & Bock, R. (2006). Constancy of organellar genome copy numbers during leaf development and senescence in higher plants. *Molecular Genetics and Genomics*, 275(1), 185-192.
- Li, X.-Q., Chetrit, P., Mathieu, C., Vedel, F., De Paepe, R., Remy, R., & Ambard-Bretteville, F. (1988). Regeneration of cytoplasmic male sterility protoclones of *Nicotiana sylvestris* with mitochondrial variations. *Current Genetics*, 13(1), 261-266.
- Li, Y., Zhang, Z., Yang, J., & Lv, G. (2018). Complete chloroplast genome of seven *Fritillaria* species, variable DNA markers identification and phylogenetic relationships within the genus. *PLoS ONE*, 13(3), e0194613.
- Librado, P., & Rozas, J. (2009). DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, 25(1), 1451-1452.
- Lim, E. T. K. (1991). *A comparative study on the performance of Metroxylon sagu and M. rumphii grown on gleyed mineral oil and organic soil* (Master's thesis). Universiti Pertanian Malaysia, Selangor, Malaysia.

- Lim, L. W. K., Chung, H. H., Chong, Y. L., & Lee, N. K. (2018). A survey of recently emerged genome-wide computational enhancer predictor tools. *Computational Biology and Chemistry*, 74(1), 132-141.
- Lim, L. W. K., Chung, H. H., & Hussain, H. (2020). Complete chloroplast genome sequencing of sago palm (*Metroxylon sagu* Rottb.): Molecular structures, comparative analysis and evolutionary significance. *Gene Rep.*, 19(1), 100662.
- Lim, L. W. K., Lau, M. L. L., Chung, H. H., Hasnain, H., & Gan, H. M. (2022). First high-quality genome assembly data of sago palm (*Metroxylon sagu* Rottboll). *Data in Brief*, 40(1), 107800.
- Lin, C. P., Huang, J. P., Wu, C. S., Hsu, C. Y., & Chaw, S. M. (2010). Comparative chloroplast genomics reveals the evolution of Pinaceae genera and subfamilies. *Genome Biology and Evolution*, 2(1), 504-517.
- Linggang, S., Phang, L. Y., Wasoh, M. H., & Abd-Aziz, S. (2012). Sago pith residue as an alternative cheap substrate for fermentable sugars production. *Applied Biochemistry and Biotechnology*, 167(1), 122-131.
- Little, D. Y., Rao, H., Olive, S., Dabiel-Vedele, F., Krapp, A., & Malamy, J. E. (2005). The putative high-affinity nitrate transporter NRT2.1 represses lateral root initiation in response to nutritional cues. *Proceedings of the National Academy of Sciences*, 102(1), 13693–13698.
- Liu, N., Xue, Y., Guo, Z., Li, W., & Tang, J. (2016). Genome-wide association study identifies candidate genes for starch content regulation in maize kernels. *Frontiers in Plant Science*, 7(1), 1046.

- Liu, P., & Demple, B. (2010). DNA repair in mammalian mitochondria: Much more than we thought? *Environmental and Molecular Mutagenesis*, 51(1), 417-426.
- Liu, X., Li, Y., Yang, H., & Zhou, B. (2018). Chloroplast genome of folk medicine and vegetable plant *Talinum paniculatum* (Jacq.) Gaertn: Gene organization, comparative and phylogenetic analysis. *Molecules*, 23(4), 857.
- Lohse, M., Drechsel, O., Kahlau, S., & Bock, R. (2013). Organellar genome DRAW: A suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Research*, 41(1), W575-W581.
- Lowe, T. M., & Chan, P. P. (2016). tRNAscan-SE on-line: Search and contextual analysis of transfer RNA genes. *Nucleic Acids Research*, 44(1), W54-W57.
- Ma, J., & Li, X.-Q. (2015). Organellar genome copy number variation and integrity during moderate maturation of roots and leaves of maize seedlings. *Current Genetics*, 61(4), 591-600.
- Mantere, T., Kersten, S., & Hoischen, A. (2019). Long-read sequencing emerging in medical genetics. *Frontiers in Genetics*, <https://doi.org/10.3389/fgene.2019.00426>.
- Mardanov, A. V., Ravin, N. V., Kuznetsov, B. B., Samigullin, T. H., Antonov, A. S., Kolganova, T. V., & Skyabin, K. G. (2008). Complete sequence of the duckweed (*Lemna minor*) chloroplast genome: Structural organization and phylogenetic relationships to other angiosperms. *Journal of Molecular Evolution*, 66(1), 555-564.
- Matoh, T. (2002). Salt stress. In *Encyclopedia of plant nutrition and fertilizer* (pp. 319-321). Tokyo, Japan: Asakura Shoten.

- Mayonde, S., Cron, G. V., Gaskin, J. F., & Byrne, M. J. (2015). Evidence of *Tamarix* hybrids in South Africa, as inferred by nuclear ITS and plastid *trnS-trnG* DNA sequences. *South African Journal of Botany*, 96(1), 122-131.
- McFadden, G. I. (2001). Chloroplast origin and integration. *Plant Physiology*, 25(1), 50-53.
- Mel, M., Nadzri, H., Kamarudin, M. H., & Karim, M. I. A. (2010). Bioethanol production from sago starch. In Z. Ahmad (Ed.), *Sago (Metroxylon Rottb) and its applications*. Kuala Lumpur, Malaysia: IIUM Press.
- Ming, R., VanBuren, R., Liu, Y., Yang, M., Han, Y., Li, L. T., Zhang, Q., Kim, M. J., Schatz, M. C., Campbell, M., Li, J., Bowers, J. E., Tang, H., Lyons, E., Ferguson, A. A., Narzisi, G., Nelson, D. R., Blaby-Haas, C. E., Gschwend, A. R., Jiao, Y., ... Shen-Shen-Miller, J. (2013). Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). *Genome Biology*, 14(1), R41.
- Ming, R., VanBuren, R., Wai, C. M., Tang, H., Schatz, M. C., Bowers, J. E., Lyons, E., Wang, M. L., Chen, J., Biggers, E., Zhang, J., Huang, L., Zhang, L., Miao, W., Zhang, J., Ye, Z., Miao, C., Lin, Z., Wang, H., ... Yu, Q., (2015). The pineapple genome and the evolution of CAM photosynthesis. *Nature Genetics*, 47(1), 1435–1442.
- Mizuno, K., Kobayashi, E., Tachibana, M., Kawasaki, T., Fujimura, T., Funane, K., Kobayashi, M., & Baba, T. (2001). Characterization of an Isoform of Rice starch branching enzyme, RBE4, in developing seeds. *Plant Cell Physiology*, 42(1), 349–357.

- Monib, N. J. (2015). *Development on sago sugars recovery and purification by filtration on powdered activated carbon (PAC)* (Master's thesis). Universiti Malaysia Sarawak, Sarawak, Malaysia.
- Muljana, H., Sugih, A. K., Christina, N., Fangdinata, K., Renaldo, J., Rudy, Heeres, H. J., & Picchioni, F. (2016). Transesterification of sago starch and waste palm cooking oil in densified CO₂. *IOP Conference Series: Materials Science and Engineering*, 223(1), 012055.
- Nakamura, T., & Takabe, T. (1999). Glycinebetaine in plants. In *Environmental response in plants: Survival strategy and its molecular mechanism* (pp. 86-90). Tokyo, Japan: Shujunsya.
- Nakamura, Y., Utsumi, Y., Sawada, T., Aihara, S., Utsumi, C., Yoshida, M., & Kitamura, S. (2010). Characterization of the reactions of starch branching enzymes from rice endosperm. *Plant and Cell Physiology*, 51(5), 776-794.
- Nisar, M., & Hussain, H. (2020). The protein extraction method of *Metroxylon sagu* leaf for high-resolution two-dimensional gel electrophoresis and comparative proteomics. *Chemical and Biology Technologies in Agriculture*, 7(1), 14.
- Nishimura, Y. (2008). Present sago palm situation in Mindanao, Philippines. In *Proceedings of the 17th conference of the society of sago palm studies* (pp. 31-36). Tokyo, Japan: Japanese Society Sago Palm Studies.
- Niyong, L. (2012). *Continuous bioethanol production from fermentation of hydrolysed sago starch using the 'Ishizaki Process'* [Master's thesis, Universiti Malaysia Sarawak]. Universiti Malaysia Sarawak, Sarawak, Malaysia.

- Negrak, V. (2013). Mitochondrial genome sequence of the legume *Vicia faba*. *Frontiers in Plant Science*, doi: 10.3389/fpls.2013.00128
- Notohadiprawiro, T., & Louhenapessy, J. E. (1992). Potensi sagu dalam penganeekaragaman bahan pangan pokok ditinjau dari persyaratan lahan. [The potential of sago in diversification of staple food in terms of land requirements.] In *Prosiding symposium sagu nasional* (pp. 99-106). Fakultas Pertanian Universitas Pattimura, Indonesia.
- Novero, A. U., Mabras, M. B., & Esteban, H. J. (2012). Epigenetic inheritance of spine formation in sago palm (*Metroxylon sagu* Roettb.). *Plant Omics Journal*, 5(6), 559-566.
- Nozaki, K., Nuyim, T., Shinano, T., Hamada, S., Ito, H., Matsui, H., & Osaki, M. (2004). Starch properties of the sago palm (*Metroxylon sagu* Rottb.) in different soils. *Plant Foods for Human Nutrition*, 59(1), 85-92.
- Nuttanan, S., Wichan, K., Sansanee, P., & Ampol, M. (1995). A preliminary study of sago starch as a tablet binder. *Warasan Phesatchasat*, 22(3), 101–106.
- Nystedt, B., Street, N. S., Wetterbom, A., Zuccolo, A., Lin, Y. C., Scofield, D. G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., Vicedomini, R., Sahlin, K., Sherwood, E., Elfstrand, M., Gramzow, L., Holmberg, K., Hallman, J., Keech, O., Klasson, L., ... Jansson, S. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature*, 497(1), 579-584.
- Oda, K., Yamato, K., Ohta, E., Nakamura, Y., Takemura, M., Nozato, N., Akashi, K., Kanegae, T., Ogura, Y., & Kohchi, T. (1992). Gene organization deduced from the complete sequence of liverwort *Marchantia polymorpha* mitochondrial DNA, A

- primitive form of plant mitochondrial genome. *Journal of Molecular Biology*, 223(1), 1-7.
- Ogita, S., Kubo, T., Takeuchi, M., Yamaguchi, C., & Okazaki, M. (1996). Accumulations and distribution of starch in sago palm (*Metroxylon sagu*) stems. *Sago Palm*, 4(1), 1-5.
- Ohdan, T., Francisco, P. B., Sawada, T., Hirose, T., Terao, T., Satoh, H., & Nakamura, Y. (2005). Expression profiling of genes involved in starch synthesis in sink and source organs of rice. *Journal of Experimental Botany*, 56(422), 3229-3244.
- Ohyama, K., Fukuzawa, H., Kohchi, T., Shirai, H., Sano, T., Sano, S., Umesono, K., Shiki, Y., Takeuchi, M., Chang, Z., Aota, S. I., Inokuchi, H., & Ozeki, H. (1986). Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. *Nature*, 322(1), 572-574.
- Okazaki, M., & Sasaki, Y. (2018). Soil environment in sago palm forest. In H. Ehara, Y. Toyoda, & D. Johnson (Eds.), *Sago palm: Multiple contributions to food security and sustainable livelihoods* (pp. 193-206). Singapore: Springer.
- Oldenburg, D. J., Rowan, B. A., Kumar, R. A., & Bendich, A. J. (2014). On the fate of plastid DNA molecules during leaf development: Response to the Golczyk *et al.* commentary. *Plant Cell*, 26(1), 855-861.
- Oldenburg, D. J., Rowan, B. A., Zhao, L., Walcher, C. L., Schleh, M., & Bendich, A. J. (2006). Loss or retention of chloroplast DNA in maize seedlings is affected by both light and genotype. *Planta*, 225(1), 41-55.
- Palmer, J. D., & Delwiche, C. F. (1998). The origin and evolution of plastids and their genomes. In D. E. Soltis, P. S. Soltis, J. J. Doyle, M. A. Norwell (Eds.), *Molecular systematics of plants, II, DNA sequencing*. Netherlands: Kluwer Academic.

- Park, I., Yang, S., Choi, G., Kim, W. J., & Moon, B. C. (2017). The complete chloroplast genome sequences of *Aconitum pseudolaeve* and *Aconitum longecassidatum*, and development of molecular markers for distinguishing species in the *Aconitum* subgenus *Lycoctonum*. *Molecules*, 22(11), e2012.
- Park, I., Yang, S., Kim, W. K., Song, J. H., Lee, H. S., Lee, H. O., Lee, J. H., Ahn, S. N., & Moon, B. C. (2019). Sequencing and comparative analysis of the chloroplast genome of *Angelica polymorpha* and the development of a novel indel marker for species identification. *Molecules*, 24(6), 1038.
- Patankar, H. V., Al-Harrasi, I., Al-Tahyai, R., & Yaish, M. W. (2018). Identification of candidate genes involved in the salt tolerance of date palm (*Phoenix dactylifera* L.) based on a yeast functional bioassay. *DNA and Cell Biology*, 37(6), 524-534.
- Patankar, H. V., Al-Harrasi, I., Al-Tahyai, R., & Yaish, M. W. (2019). Functional characterization of date palm aquaporin gene PdPIP1;2 confers drought and salinity tolerance to yeast and Arabidopsis. *Genes (Basel)*, 10(5), 390.
- Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., Schmutz, J., Spannagl, M., Tang, H., Wang, X., Wicker, T., Bharti, A. K., Chapman, J., Feltus, F. A., Gowik, U., ... Rokhsar, D. S. (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, 457(1), 551-556.
- Pei-Lang, A. T., Mohamed, A. M. D., & Karim, A. A. (2006). Sago starch and composition of associated components in palms of different growth stages. *Carbohydrate Polymers*, 63(2), 283-286.

- Pellicer, J., Fay, M. F., & Leitch, I. J. (2010). The largest eukaryotic genome of them all. *Botanical Journal of Linnean Society*, 164(1), 10-15.
- Petersen, T.N., Brunak, S., von Heijne, G. & Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, 8(1), 785-786.
- Pfister, B., & Zeeman, S. C. (2016). Formation of starch in plant cells. *Cellular and Molecular Life Sciences*, 73(1), 2781-2807.
- Pickett, B. D., Miller, J. B., & Ridge, P. G. (2017). Kmer-SSR: A fast and exhaustive SSR search algorithm. *Bioinformatics*, 33(24), 3922-3928.
- Prathumyot, W., Okada, M., Naito, H., & Ehara, H. (2011). Physiological response and mineral concentration of sago palm diurnal changes of NaCl concentration in culture solution. *Tropical Agriculture Development*, 55(1), 11-20.
- Preuten, T., Cincu, E., Fuchs, J., Zoschke, R., Liere, K., & Borner, T. (2010). Fewer genes than organelles: extremely low and variable gene copy numbers in mitochondria of somatic plant cells. *Plant Journal*, 64(1), 948-959.
- Priya, M., Meenambal, T., Balasubramanian, N., & Perumal, B. (2015). Comparative study of treatment of sago wastewater using HUASB reactor in the presence and absence of effective microorganisms. *Procedia Earth and Planetary Science*, 11(1), 483-490.
- Purwani, E. Y., Widaningrum Thahir, R., & Muslich (2006). Effect of heat moisture treatment of sago starch on its noodle quality. *Indonesian Journal of Agricultural Science*, 7(1), 8-14.

- Purwoko, D., Cartealy, I. C., Tajuddin, T., Dinarti, D., & Sudarsono, S. (2019). SSR identification and marker development for sago palm based on NGS genome data. *Breeding Science*, 69(1), 1-10.
- Quat Ng, N. (2007). Guest speech by FAO at the opening session. In M. A. Quevedo & A. B. Loreto, *Abstracts of the Ninth International Sago Symposium*. The Philippines: Ormoc City.
- Rauwerdink, J. B. (1986). An essay on *Metroxylon*, the sago palm. *Principes: Journal of the International Palm Society*, 30(4), 165-180.
- Rawung, J. B. M., & Indrasti, R. (2021). The constraints to sago development and improvement efforts in Siau Tagulandang Biaro (Sitaro) Islands. *E3S Web of Conference*, 232, 01029.
- Radley, J. A. (1976). The textile industry. In J. A. Radley, *Industrial uses of starch and its derivatives* (pp. 149–198). London: Applied Science Publishers Ltd.
- Ravi, V., Khurana, J. P., Tyagi, A. K., & Khurana, P. (2007). An update on chloroplast genomes. *Plant Systematics and Evolution*, 271(1), 101-122.
- Rayner, J. C., & Pelham, H. R. (1997). Transmembrane domain-dependent sorting of proteins to the ER and plasma membrane in yeast. *EMBO Journal*, 16(8), 1832-1841.
- Regina, A., Bird, A., Topping, D., Bowden, S., Freeman, J., Barsby, T., Kosar-Hashemi, B., Li, Z., Rahman, S., & Morell, M. (2006). High-amylose wheat generated by RNA interference improves indices of large-bowel health in rats. *The Proceedings of the National Academy of Sciences*, 103(1), 3546–3551.

- Richard, G. F., Kerrest, A., & Dujon, B. (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and Molecular Biology Reviews*, 72(4), 686-727.
- Rijkers, T., Pons, T. L., & Bongers, F. (2001). The effect of tree height and light availability on photosynthetic leaf traits of four neotropical species differing in shade tolerance. *Functional Ecology*, 14(1), 77-86.
- Roslan, H., Hossain, M. A., & Gerunsin, J. (2017). Molecular and 3D-structural characterization of fructose-1,6-bisphosphate aldolase derived from *Metroxylon sagu*. *Brazilian Archives of Biology and Technology*, 60(1), e17160108.
- Roslan, H. A. (2012). Sago research in UNIMAS Sarawak: Gene prospecting from sago genome. In N. Busri, Z. C. Abdullah, C. N. Hipolito, H. A. Roslan, E. U. H., Sim, D. S. Awg-Adeni, A. A. S. A. Husaini, & N Humrawali (Eds.), *Proceedings of 2nd ASEAN Sago Symposium 2012: Advances in Sago Research and Development* (p. 16). Sarawak, Malaysia: Universiti Malaysia Sarawak.
- Roslan, H. A., & Anji, S. B. (2011). Characterization of inflorescence-predominant *chitinase* gene in *Metroxylon sagu* via differential display. *3 Biotech*, 1(1), 27-33.
- Rowan, B. A., & Bendich, A. J. (2009). The loss of DNA from chloroplasts as leaves mature: Fact or artefact? *Journal of Experimental Botany*, 60(1), 3005-3010.
- Ruddle, K. R. (1976). Sago in the new world. In Tan K. (Ed.), *Sago-76, Proc 1st Sago Symp*, Sarawak, Malaysia.
- Saithong, T., Rongsirikul, O., Kalapanulak, S., Chiewchankaset, P., Siriwat, W., Netrphan, S., Suksangpanomrung, M., Meechai, A., & Cheevadhanarak, S. (2013). Starch

biosynthesis in cassava: A genome-based pathway reconstruction and its exploitation in data integration. *BMC Systems Biology*, 7(1), 75.

Sangeetha, V., & Sivakumar, V. (2016). Biogas production from synthetic sago wastewater by anaerobic digestion: Optimization and treatment. *Chemical Industry & Chemical Engineering Quarterly*, 22(1), 33-39.

Saravanane, R., Murthy, D. V., & Krishnaiah, K. (2001). Anaerobic treatment and biogas recovery for sago wastewater management using a fluidized bed reactor. *Water Science & Technology*, 44(6), 141-146.

Sasaoka, M., Laumonier, Y., & Sugimura, K. (2014). Influence of indigenous sago-based agriculture on local forest landscapes in Maluku, East Indonesia. *Journal of Tropical Forest Science*, 26(1), 75-83.

Sawada, T., Itoh, M., & Nakamura, Y. (2018). Contributions of three starch branching enzyme isozymes to the fine structure of amylopectin in rice endosperm. *Frontiers in Plant Science*, 9(1), 1536.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F. S., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., ... Wilson, R. K. (2009). The B73 maize genome: Complexity, diversity, and dynamics. *Science*, 326(5956), 1112-1115.

Schmitz-Linnerweber, C., & Barkan, A. (2007). RNA splicing and RNA editing in chloroplasts. *Topics in Current Genetics*, 19(1), 213-248.

- Schuiling, D. L. (2009). *Growth and development of true sago palm (Metroxylon sagu Rottboll.) with special reference to accumulation of starch in the trunk* (Doctoral dissertation). Wageningen University, Wageningen.
- Schuiling, D. L., & Flach, M. (1985). *Guidelines for the cultivation of sago palm*. Agricultural University, Wageningen: Department of Tropical Crop Science.
- Segui-Simarro, J. M., Coronado, M. J., & Staehelin, L. A. (2008). The mitochondrial cycle of Arabidopsis shoot apical meristem and leaf primordium meristematic cells is defined by a perinuclear tentaculate/cage-like mitochondrion. *Plant Physiology*, 148(1), 1380-1393.
- Shahid Masood, M., Nishikawa, T., Fukuoka, S., Njenga, P. K., Tsudzuki, T., & Kadowaki, K. (2004). The complete nucleotide sequence of wild rice (*Oryza nivara*) chloroplast genome: First genome wide comparative sequence analysis of wild and cultivated rice. *Gene*, 340(1), 133-139.
- Shangguan, L. F., Han, J., Kayesh, E., Sun, X., Zhang, C. Q., Pervaiz, T., Wen, X., & Fang, J. (2013). Evaluation of genome sequencing in selected plant species using expressed sequence tags. *PLoS ONE*, 8(1), e69890.
- Shaver, J. M., Oldenburg, D. J., & Bendich, A. J. (2006). Changes in chloroplast DNA during development in tobacco, *Medicago truncatula*, pea, and maize. *Planta*, 224(1), 72-82.
- Shi, M. J., Chen, Y. Y., Zhang, W. T., & Xia, X. Q. (2016). The evolutionary mechanism of genome size. *Chinese Science Bulletin*, 61(1), 3188-3195.

- Shibata, K., Flores, D. M., Kobayashi, G., & Sonomoto, K. (2007). Direct L-lactic acid fermentation with sago starch by a novel amylolytic lactic acid bacterium, *Enterococcus faecium*. *Enzyme and Microbial Technology*, 41(1-2), 149-155.
- Shima, A. R. R., Salina, H. F., Masniza, M., & Atiqah, A. H. (2012). Viability of lactic acid bacteria in home made yogurt containing sago starch oligosaccharides. *International Journal of Basic & Applied Sciences*, 12(1), 58-62.
- Shimizu, T., Kaminuma, E., Nonaka, K., Yoshioka, T., Goto, S., Matsumoto, T., Katayose, Y., Mochizuki, T., Tanizawa, Y., Toyoda, A., Fujiyama, A., & Nakamura, Y. (2016). A genomic approach to selecting robust and versatile SNP sets from next-generation sequencing data for genome-wide association study in citrus cultivars. *Acta Horticulturae*. 1135(1), 23–32.
- Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T., Hayashida, N., Matsubayashi, T., Zaita, N., Chunwongse, J., Obokata, J., Yamaguchi-Shinozaki, K., Ohto, C., Torazawa, K., Meng, B. Y., Sugita, M., Deno, H., Kamogashira, T., Yamada, K., Kusuda, J., Takaiwa, F., Kato, A., Tohdoh, N., Shimada, H., & Sugiura, M. (1986). The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *EMBO Journal*, 5(9), 2043-2049.
- Shintani, D., & DellaPenna, D. (1998). Elevating the vitamin E content of plants through metabolic engineering. *Science*, 282(1), 2098-2100.
- Shipton, W. A., Baker, A., Blaney, B. J., Horwood, P. F., Warner, J. M., Pelowa, D., & Greenhill, A. R. (2010). Nitrogen fixation associated with sago (*Metroxylon sagu*) and some implications. *Letters in Applied Microbiology*, 52(1), 56-61.

- Shrestha, A., Toyota, K., Nakano, Y., Okazaki, M., Quevedo, M. A., Loreto, A. B., Mariscal, A., & Abayon, E. I. (2006). Nitrogen-fixing activity in various parts of sago palm (*Metroxylon sagu*) and characterization of aerobic nitrogen-fixing bacteria colonizing in sago palm. *Sago Palm*, 14(1), 20-32.
- Shrestha, A., Toyota, K., Okazaki, M., Suga, Y., Quevedo, M. A., Loreto, A. B., & Mariscal, A. A. (2007). Enhancement of nitrogen-fixing activity of Enterobacteriaceae strains isolated from sago palm (*Metroxylon sagu*) by microbial interaction with non-nitrogen fixers. *Microbes and Environments*, 22(1), 59-70.
- Silva, G. A. R., Jojima, C., Moraes, E. M., Antonelli, A., Manfrin, M. H., & Franco, F. F. (2016). Intra and interspecific sequence variation in closely related species of *Cereus* (CACTACEAE). *Biochemical Systematics and Ecology*, 65(1), 137-142.
- Sim, E. S., & Ahmad, M. I. (1991). Leaf nutrient variations in sago palms. In T. T. Ng, Y. L. Tie, & H. S. Kueh (Eds.), *Proceedings of the fourth international sago symposium* (pp. 92-102). Sarawak, Malaysia: Lee Ming Press.
- Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy ortholog. *Bioinformatics*, 31(19), 3210-3212.
- Singh, R., Ong-Abdullah, M., Low, E. T. L., Manaf, M. A. A., Rosli, R., Nookiah, R., Ooi, L. C. L., Chan, K. L., Halim, M. A., Azizi, N., Nagappan, J., Bacher, B., Lakey, N., Smith, S. W., He, D., Hogan, M., Budiman, M. A., Lee, E. K., DeSalle, R., ... Sambanthamurthi, R. (2013). Oil palm genome sequence reveals divergence of interfertile species in old and new worlds. *Nature*, 500(1), 335-339.

- Singhal, R. S., Kennedy, J. F., Gopalakrishnan, S. M., Kaczmarek, A., Knill, C. J., & Akmar, P. F. (2008). Industrial production, processing, and utilization of sago palm-derived products. *Carbohydrate Polymers*, 72(1), 1-20.
- Sloan, D. B., Alverson, A. J., Chuckalovcak, J. P., Wu, M., McCauley, D. E., Palmer, J. D., & Taylor, D. R. (2012). Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS Biology*, 10(1), e1001241.
- Smith, K. (2015). Cultivation and management. In K. Smith (Ed.), *The sago palm: The food and environmental challenges of the 21st century*. Kyoto, Japan: Kyoto University Press.
- Song, Y., Dong, W., Liu, B., Xu, C., Yao, X., Gao, J., & Corlett, R. T. (2015). Comparative analysis of complete chloroplast genome sequences of two tropical trees *Machilus yunnanensis* and *Machilus balansae* in the family Lauraceae. *Frontier in Plant Sciences*, 6(1), 662.
- Su, H. J., Hogenhout, S. A., Al-Sadi, A. M., & Kuo, C. H. (2014). Complete chloroplast genome sequence of Omani lime (*Citrus aurantiifolia*) and comparative analysis within the rosids. *PLoS ONE*, 9, e113049.
- Su, J., Zhu, L., Sherman, A., Wang, X., Lin, S., Kamesh, A., Norikane, J. H., Streatfield, S. J., Herzog, R. W., & Daniell, H. (2015). Low cost industrial production of coagulation factor IX bioencapsulated in lettuce cells for oral tolerance induction in haemophilia B. *Biomaterials*, 70(1), 84-93.
- Sullivan, A. R., Eldjell, Y., Schiffthaler, B., Delhomme, N., Asp, T., Hebelstrup, K. H., Keech, O., Oberg, L., Moller, I. M., Arvestad, L., Street, N. R., & Wang, X. R. (2020).

- The mitogenome of Norway spruce and a reappraisal of mitochondrial recombination in plants. *Genome Biology and Evolution*, 12(1), 3586-3598.
- Sumardiono, S., Jos, B., Pudjihasturi, I, Yafiz, A. M., Rachmasari, M., & Cahyano, H. (2021). Physiochemical properties of sago ozone oxidation: The effect of reaction time, acidity, and concentration of starch. *Foods*, 10(1), 1309.
- Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., Vermet, T., Corthier, G., Brochmann, C., & Willersley. (2007). Power and limitations of the chloroplast trnL(UAA) intron for plant DNA barcoding. *Nucleic Acids Research*, 35(3), e14.
- Tadano, T. (1986). Genetic resources and biotechnology in improvement of stress tolerance of crop plants, 3-b, physiological mechanisms of salt tolerance in crop plants. *Agricultural Technology*, 41(1), 337-342.
- Takanashi, S., & Hirao, K. (1992). Food-cultural studies on cooking and processing properties of sago starch. *Bulletin of the Faculty of Home Economics, Kyoritsu Women's University*, 38(1), 17-23.
- Tang, Q., Ma, X. J., Mo, C. M., Pan, L. M., Wei, R. C., & Zhao, H. (2015). Genome survey analysis in *Siraitia grosvenorii*. *Guihaia*, 35(1), 786791.
- Tawaika, F., Oono, K., & Sugiura, M. (1984). The complete nucleotide sequence of a rice 17S rRNA gene. *Nucleic Acids Research*, 12(13), 5441-5448.
- Takaiwa, F., Oono, K., Iida, Y., & Sugiura M. (1985a). The complete nucleotide sequence of a rice 25S rRNA gene. *Gene*, 37(1), 255- 259.
- Takaiwa F, Oono K, & Sugiura M. (1985b). Nucleotide sequence of the 17S--25S spacer region from rice rDNA. *Plant Molecular Biology*, 4(1), 355-364

- Tanaka, Y., Dyer, T. A., & Brownlee, G. G. (1980). An improved direct RNA sequence method; its application to *Vicia faba* 5.8S ribosomal RNA. *Nucleic Acids Research*, 8(1), 1259-1272.
- Tetlow, I. J., & Emes, M. J. (2014). A review of starch-branching enzymes and their role in amylopectin biosynthesis. *International Union of Biochemistry and Molecular Biology*, 66(8), 546-558.
- Tian, N., Han, L., Chen, C., & Wang, Z. (2018). The complete chloroplast genome sequence of *Epipremnum aureum* and its comparative analysis among eight Araceae species. *PLoS ONE*, 13(3), e0192956.
- Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., & Greiner, S. (2017). GeSeq: Versatile and accurate annotation of organelle genomes. *Nucleic Acids Research*, 45(1), W6-W11.
- Toyota, K. (2018). Microbial interactions and activities affecting sago palm growth. In H. Ehara, Y. Toyoda, & D. Johnson (Eds.), *Sago Palm* (pp. 207-217). Singapore, Singapore: Springer.
- Tsoumani, K. T., & Mathiopoulos, K. D. (2012). Genome size estimation with quantitative real-time PCR in two Tephritidae species: *Ceratitis capitata* and *Bactrocera oleae*. *Journal of Applied Entomology*, 136(1), 626-631.
- Tymms, M. J., Scott, N. S., & Possingham, J. V. (1983). DNA content of *Beta vulgaris* chloroplasts during leaf cell expansion. *Plant Physiology*, 71(4), 785-788.
- Udy, D. B., Belcher, S., Williams-Carrier, R., Gualberto, J. M., & Barkan, A. (2012). Effects of reduced chloroplast gene copy number on chloroplast gene expression in maize. *Plant Physiology*, 160(3), 1420-1431.

- Uthaipaisanwong, P., Chanprasert, J., Shearman, J. R., Sangsrakru, D., Yoocha, T., Jomchai, N., Jantasuriyarat, C., Tragoonrung, S., & Tangphatsornruang, S. (2012). Characterization of the chloroplast genome sequence of oil palm (*Elaeis guineensis* Jacq.). *Gene*, 500(2), 172-180.
- van Dingenan, J., Blomme, J., Gonzalez, N., & Inzé, D. (2016). Plants grow with a little help from their organelle friends. *Journal of Experimental Botany*, 67(22), 6267-6281.
- Vennilamani, N., Kadirvelu, K., Sameena, Y., & Pattabhi, S. (2005). Utilization of activated carbon prepared from industrial solid waste for the removal of chromium (VI) ions from synthetic solution and industrial effluent. *Adsorption Science and Technology*, 23(2), 145–160.
- Verma, D., Kanagaraj, A., Jin, S. X., Singh, N. D., Kolattukudy, P. E., & Daniell, H. (2010). Chloroplast-derived enzyme cocktails hydrolyse lignocellulosic biomass and release fermentable sugars. *Plant Biotechnology Journal*, 8(1), 332-350.
- Vickineswary, S., Shim, Y. L., Thambirajah, J. J., & Blakebrough, N. (1994). Possible microbial utilisation of sago processing wastes. *Resource and Conservation Recycling*, 11(1), 289–296.
- Vickineswary, S., Getha, K., Maheswary, S., Chong, V. C., Shaliza, I., & Sastry, C. A. (1997). Growth of *Rhodopseudomonas palustris* strain B1 in sago starch processing waste water. *Studies in Environmental Sciences*, 66(1), 335-348.
- Vothknecht, U. C., & Westhoff, P. (2001). Biogenesis and origin of thylakoid membranes. *Biochimica et Biophysica Acta*, 1541(1), 91-101.

- Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J. G., & Scharz, M. C. (2017). GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics*, 33(14), 2202-2204.
- Wada, K. (1999). Mechanism of salt tolerance. *Heredity*, 53, 58-62.
- Wai, T., Teoli, D., & Shoubridge, E. A. (2008). The mitochondrial DNA genetic bottleneck results from replication of a subpopulation of genomes. *Nature Genetics*, 40, 1484-1488.
- Wang, C. L., Ding, M. Q., Zou, C. Y., Zhu, X. M., Tang, Y., Zhou, M. L., & Shao, J. R. (2017). Comparative analysis of four buckwheat species based on morphology and complete chloroplast genome sequences. *Scientific Reports*, 7(1), 6514.
- Wang, D. Y., Zhang, Q., Liu, Y., Lin, Z. F., Zhang, S. X., Sun, M. X., & Sodmergen. (2010). The levels of male gametic mitochondrial DNA are highly regulated in angiosperms with regard to mitochondrial inheritance. *Plant Cell*, 22(1), 2402-2416.
- Wang, J., Hu, P., Lin, L., Chen, Z., Liu, Q., & Wei, C. (2018). Gradually decreasing starch branching enzyme expression is responsible for the formation of heterogeneous starch granules. *Plant Physiology*, 176(1), 582-595.
- War, A. R., Paulraj, M. G., Ahmad, T., Buhroo, A. A., Hussain, B., Ignacimuthu, S., & Sharma, H. C. (2012). Mechanisms of plant defense against insect herbivores. *Plant Signaling & Behavior*, 7(10), 1306-1320.
- Weber, B., Zicola, J., Oka, R., & Stam, M. (2016). Plant enhancers: A call for discovery. *Trends in Plant Science*, 21(11), 974-987.
- Wee, C. C., & Roslan, H. A. (2012a). Expressed sequence tags (ESTs) from young leaves of *Metroxylon sagu*. *3 Biotech*, 2(3), 211-218.

- Wee, C. C., & Roslan, H. A. (2012b). Isolation of alcohol dehydrogenase cDNA and basal regulatory region from *Metroxylon sagu*. *ISRN Molecular Biology*, 839427(1), 1-9.
- Wilhelm, J., Pingoud, A., Hahn, M., 2003. Real-time PCR based method for the estimation of genome size. *Nucleic Acids Research*, 31(1), e56.
- Wolfe, K. H., Li, W. H., & Sharp, P. M. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *The Proceedings of the National Academy of Sciences*, 84(1), 9054-9058.
- Won, S. Y., Jung, J. A., & Kim, J. S. (2018). The complete mitochondrial genome sequence of *Chrysanthemum boreale* (Asteraceae). *Mitochondrial DNA Part B*, 3(2), doi: 10.1080/23802359.2018.1468226
- Wu, L. W., Yen, S. H., Lees, D. C., & Hsu, Y. F. (2010). Elucidating genetic signatures of native and introduced populations of the Cycad Blue, *Chilades pandava* to Taiwan: A threat both to sago palm and to native *Cycas* populations worldwide. *Biological Invasions*, 12(8), 2649-2669.
- Wu, Y. F., Xiao, F. M., Xu, H. N., Zhang, T., & Jiang, X. M. (2014). Genome survey in *Cinnamomum camphora* L. Presl. *Journal of Plant Genetic Resources*, 15(1), 149152.
- Yaish, M. W., & Kumar, P. P. (2015). Salt tolerance research in date palm (*Phoenix dactylifera* L.), past, present, and future perspectives. *Frontiers in Plant Science*, 6(1), 348.
- Yaish, M. W., Patankar, H. V., Assaha, D. V. M., Zheng, Y., Al-Yahyai, R., & Sunkar, R. (2017). Genome-wide expression profiling in leaves and roots of date palm (*Phoenix dactylifera* L.) exposed to salinity. *BMC Genomics*, 18(1), 246.
- Yamamoto, Y., Yoshida, T., & Yanagidate, I. (2008). Sago palm and its utilization in Ambon and Seram Island, Maluku Province, Indonesia. In *Proceedings of the 17th*

- conference of the society of sago palm studies* (pp. 17-22). Tokyo, Japan: Japanese Society Sago Palm Studies.
- Yang, M., Zhang, X., Liu, G., Yin, Y., Chen, K., Yun, Q., Zhao, D., Al-Mssallem, I. S., & Yu, J. (2010). The complete chloroplast genome sequence of date palm (*Phoenix dactylifera* L.). *PLoS ONE*, 15(9), e12762.
- Yeo, A., Flowers, S., Rao, G., Welfare, K., Senanayake, N., & Flowers, T. (1999). Silicon reduces sodium uptake in rice (*Oryza sativa* L.) in saline conditions and this is accounted for by a reduction in the transpirational bypass flow. *Plant Cell Environment*, 22(1), 559-565.
- Yoneta, R., Okazaki, M., Toyota, K., Yano, Y., & Power, P. (2003). Glycine betaine concentrations in sago palm (*Metroxylon sagu*) under different salt stress conditions. In *Abstracts of the 12th conference of the society of sago palm studies* (pp. 18-22). Tokyo, Japan: Japanese Society Sago Palm Studies.
- Yoneta, R., Okazaki, M., Yano, Y., & Power, A. P. (2004). Possibility of osmotic pressure regulation using K⁺ in sago palm (*Metroxylon sagu*). In *Proceedings of the 13th conference of the society of sago palm studies* (pp. 17-22). Sarawak, Malaysia: The Society of Sago Palm Studies.
- Yoneta, R., Okazaki, M., & Yano, Y. (2006). Response of sago palm (*Metroxylon sagu* Rottb.) to NaCl stress. *Sago Palm*, 14(1), 10-19.
- Yu, C-S., Chen, Y.-C., Lu, C.-H. & Hwang, J.-K. (2006). Prediction of protein subcellular localization. *Proteins: Structure, Function and Bioinformatics*, 64(3), 643-651.
- Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics*, 10(1), 451-481.

- Zhang, J., Khan, S. A., Hasse, C., Ruf, S., Heckel, D. G., & Bock, R. (2015). Pest control. Full crop protection from an insect pest by expression of long double-stranded RNAs in plastids. *Science*, 347(1), 991-994.
- Zhao, Y., Williams, R., Prakesh, C. S., & He, G. (2013). Identification and characterization of gene-based SSR markers in date palm (*Phoenix dactylifera* L.). *BMC Plant Biology*, 12(1), 237.
- Zheng, Q., Oldenburg, D. J., & Bendich, A. J. (2011). Independent effects of leaf growth and light on the development of the plastid and its DNA content in *Zea* species. *Journal of Experimental Botany*, 62(1), 2715-2730.
- Zhong, Y. D., Zhang, X., Li, Y. Q., Liu, L. P., & Yu, F. X. (2017). Genome survey of *Liriodendron chinense* (Hemsl.) Sarg. *Molecular Plant Breeding*, 15(1), 507512.
- Zhu, B., Zhang, W., Zhang, T., Liu, B., & Jiang, J. (2015). Genome-wide prediction and validation of intergenic enhancers in *Arabidopsis* using open chromatin signatures. *The Plant Cell*, 27(1), 2415-2426.
- Zicola, J. (2017). *Characterization of transcriptional enhancers in plants using Arabidopsis thaliana and Zea mays as model species* [Doctoral dissertation, University of Cologne]. University of Cologne, Germany.
- Zong, D., Zhou, A., Zhang, Y., Zou, X., Li, D., Duan, A., & He, C. (2019). Characterization of the complete chloroplast genomes of five *Populus* species from the western Sichuan plateau, southwest China: Comparative and phylogenetic analyses. *PeerJ*, 7(1), e6386.

Zoschke, R., Liere, K., & Borner, T. (2007). From seedling to mature plant: *Arabidopsis* plastidial genome copy number, RNA accumulation and transcription are differentially regulated during leaf development. *Plant Journal*, 50(1), 710-722.

APPENDICES

Appendix A: Long-PCR gel image

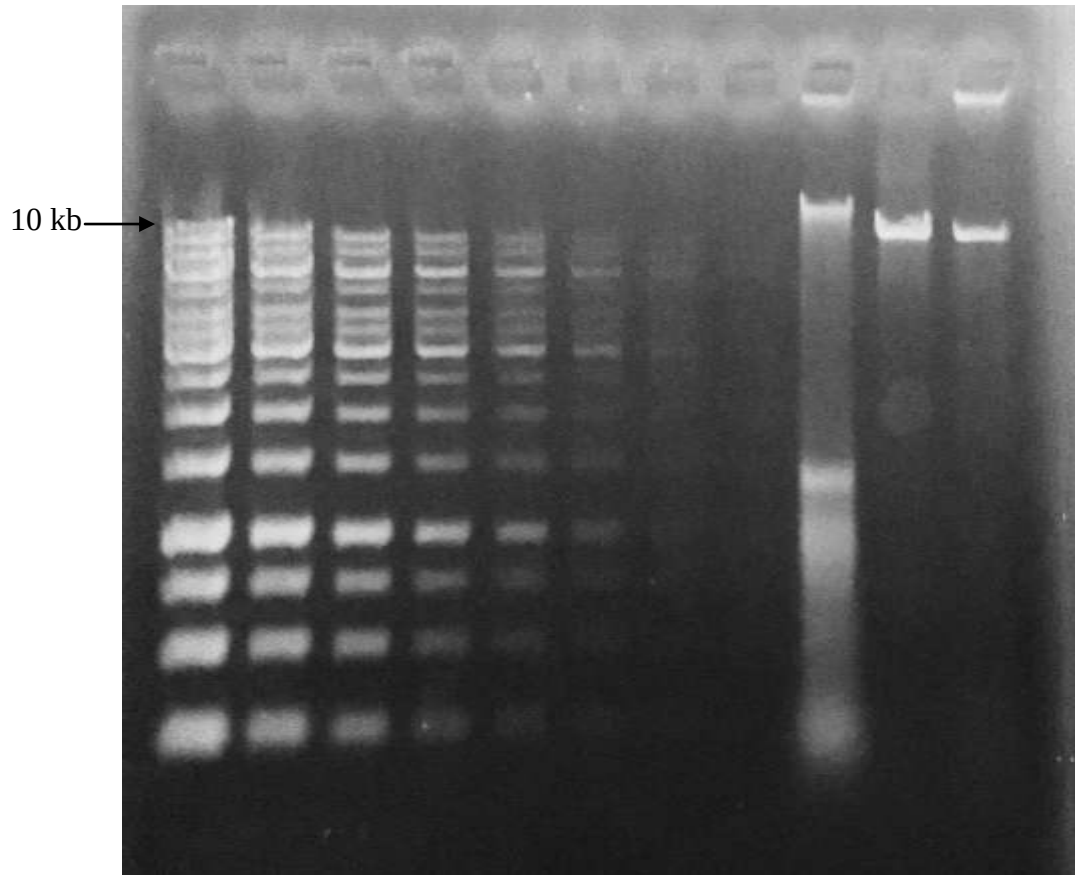


Figure 1: The long-PCR gel image. Lane 1-8: GeneRuler 1 kb DNA ladder at 1500 ng, 750 ng, 375 ng, 187.5 ng, 93.75 ng, 46.88 ng, 23.44 ng and 11.72 ng; Lane 9: genomic DNA; Lane 10: *ndh* gene; Lane 11: *rps14* gene.

Appendix B: Long-PCR and qPCR standard curves

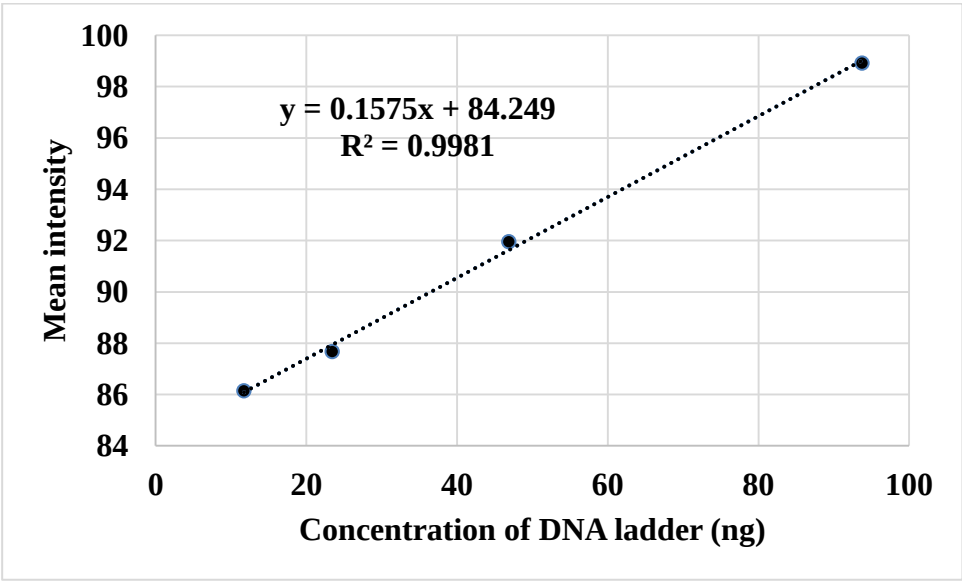


Figure 2: Long-PCR standard curve.

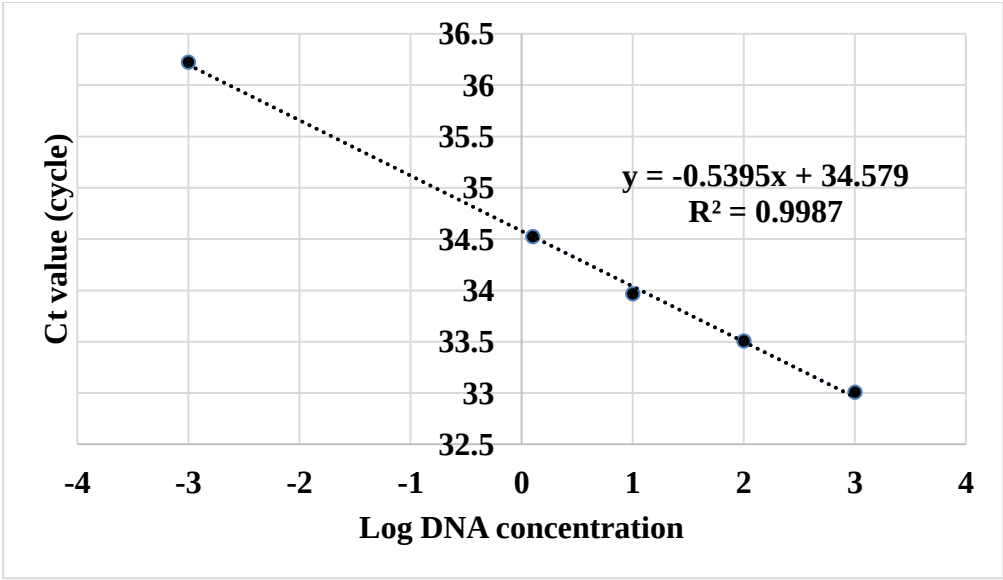


Figure 3: qPCR standard curve.

Appendix C: qPCR graphs

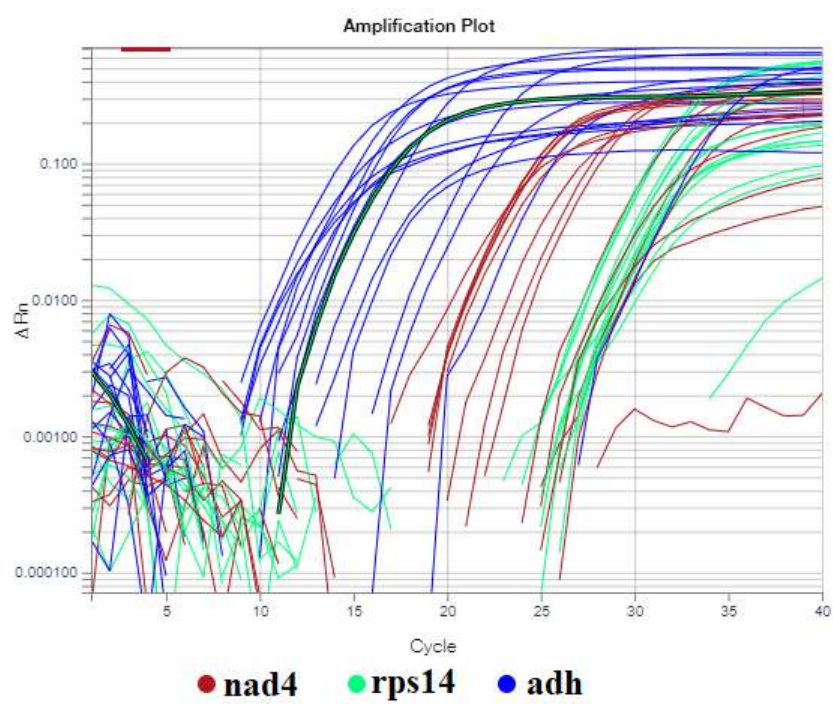


Figure 4: qPCR amplification plot.

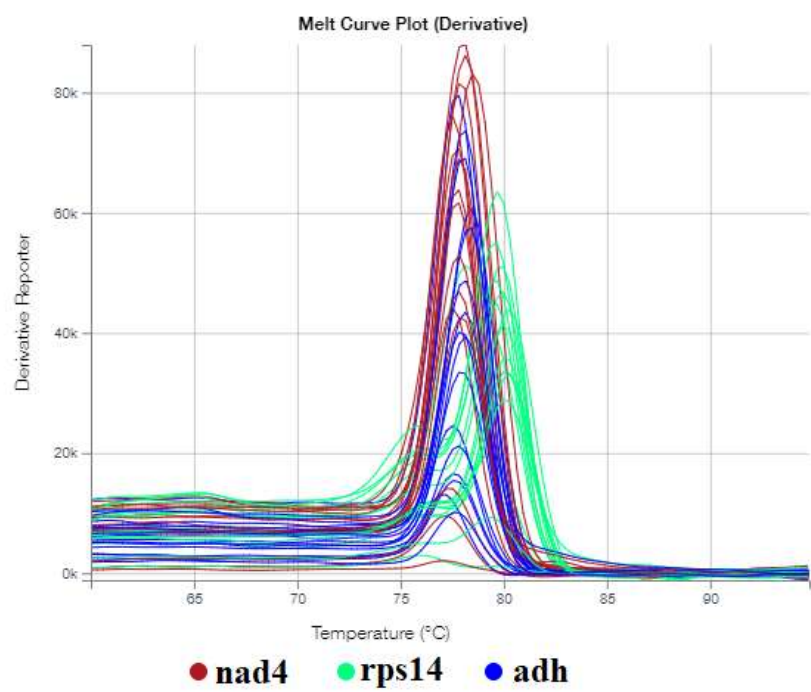


Figure 5: qPCR melt curve plot.