BIODIVERSITAS
Volume 25, Number 11, November 2024
Pages: 4043-4058

ISSN: 1412-033X
E-ISSN: 2085-4722
DOI: 10.13057/biodiv/d251104

Comparative study of chloroplast genomes across seven *Salacca* species

ARSLAN ARSHAD[1,2], REDI ADITAMA[2], MEGAYANI SRI RAHAYU[2], AZIS NATAWIJAYA[3], DEDEN DERADJAT MATRA[2], SUDARSONO SUDARSONO[1,2,♥]

[1]Program of Plant Breeding and Biotechnology, Department of Agronomy and Horticulture, Faculty of Agriculture, Institut Pertanian Bogor. Jl. Meranti, Campus IPB Dramaga, Bogor 16680, West Java, Indonesia. Tel.: +62-251-8629354, Fax.: +62-251-8629352, ♥email: sudarsono_agh@apps.ipb.ac.id

[2]PMB Lab, Department of Agronomy and Horticulture, Faculty of Agriculture, Institut Pertanian Bogor. Jl. Meranti, IPB Dramaga Campus, Bogor 16680, West Java, Indonesia

[3]Innovation Center, Corporate Development, PT. Bumitama Gunajaya Agro (Bumitama Agri Ltd.). Jl. Melawai Raya No. 10, Jakarta Selatan 12160, Jakarta, Indonesia

Manuscript received: 5 June 2024. Revision accepted: 6 November 2024.

Abstract. *Arshad A, Aditama R, Rahayu MS, Natawijaya A, Matra DD, Sudarsono S. 2024. Comparative study of chloroplast genomes across seven* Salacca *species. Biodiversitas 25: 4043-4058.* Chloroplast (Cp) genomes play a vital role in comprehending plant evolution, biodiversity, and phylogenetics. Snake fruit is a tropical fruit in the Indo-Malayan region. This work compares seven *Salacca* species Cp genomes to clarify their genetics and evolutionary connections. Cp genomes were constructed using sequencing data from the BGISeq-500 platform and the GetOrganelle assemblers. The assembled Cp genomes have a standard four-part structure and vary in length from 157,047 to 158,182 kilobase pairs (kbp). Comparative genomics analysis found the *ycf1* gene to have the highest number of single nucleotide polymorphisms (SNPs), revealing missing amino acids in *Salacca affinis.* The Cp genomes showed a high prevalence of mononucleotide SSR motifs. With a few exceptions, especially *Salacca wallichiana*, most Cp genomes showed stable borders between the large single copy (LSC), inverted repeat (IR), and short single copy (SSC) sections. This research underscores the importance of Cp genome information for identifying species, a crucial tool for evolutionary studies and breeding purposes. Furthermore, it emphasizes the intimate genetic connection between *Salacca* and *Cocos nucifera,* which contrasts with *Phoenix dactylifera.* This thorough research provides vital insights into the genetics of *Salacca* species and highlights the usefulness of Cp genome data in subsequent analyses.

Keywords: GetOrganelle, genome annotation, phylogenetic analysis, species identification

INTRODUCTION

Locally known as *salak*, snake fruit (*Salacca* sp.) is a unique tropical fruit native to Sumatra, Borneo, and the Malay Peninsula. The most extensive *Salacca* species (23 species) exist in Borneo. Most species are localized in small regions, with habitats ranging from 5 to 1700 m above sea level (ASL). *Salacca dolicholepis* is the broadest species, while *Salacca wallichiana* and *Salacca zalacca* are commonly found (Zumaidar and Miftahuddin 2018). *Salak* belongs to the Arecaceae family, which contains mangrove palm, coconut, oil palm, and betel nut fields (Ismail and Abu Bakar 2018). It is a crispy fruit with a distinct flavor that combines pineapple, banana, and apple. *Salak* is commonly grown in the lowlands of Indonesia and other Southeast Asian nations, although it is also flood- and drought-tolerant (Lestari et al. 2002). *Salacca edulis* Reinw, also known as "*salak madu*" or "honey *salak*," is a unique *salak* known for its substantial flesh, abundant water content, and delightful sweetness (Silitonga et al. 2019). One of *Salacca*'s drawbacks is its fast decomposition, which leads to waste and unpleasant smells. To resolve this problem, overripe honey *salak* that is about to perish can be turned into the profitable Nata de Salaca using a biotechnological process utilizing *Acetobacter xylinum* (Silitonga et al. 2019; Irwani et al. 2022). Overproduction of *salak* fruits may result in a significant price drop, resulting in farmers refraining from harvesting and potentially resulting in wastage. The best way to stop and solve this issue is to start producing Nata de *Salaca* (Silitonga et al. 2019). Salak peel and edible sections hold much potential for anti-inflammatory, anti-tumor, antioxidant, and anti-diabetic benefits (Saleh et al. 2018).

Even though they are economically significant, the genetic information of *Salacca* sp. is limited and needs further investigation. Understanding the chloroplast genome (Cp genome) may be the initial step to understanding *Salacca* sp. genetics. Chloroplasts are crucial for plant growth and development, facilitating carbon fixation and photosynthesis (Gan et al. 2019). Chloroplasts also contain their genome, which consists of short, circular, double-stranded DNA molecules with a size of 83-292 kb and exhibit uniparental inheritance (Ahmed 2015). A typical Cp genome comprises a pair of inverted repeat (IR) regions, separated by a large section called the long single copy (LSC) and a smaller region known as the short single copy (SSC) region (Kaila et al. 2017). The inheritance of genetic characteristics from the maternal generation in the Cp genome has provided significant and distinctive insight into plant systematics and evolutionary relationships (Wang et al. 2016). Cp genomes have been utilized for phylogenetic analysis, species identification, and population genetic studies (Zhang et al. 2016; Yu et al. 2017). A practical method for identifying plant species, particularly in taxonomically complex groups, involves the analysis of Cp genomes to provide potential markers (Chen et al. 2015; Li et al. 2015).

The Cp genome is an excellent option for identifying closely related plant species due to its short genome size, significant interspecies differences, limited intraspecies divergence, and simplicity of modification (Li et al. 2015).

The rapid progress in next-generation sequencing (NGS) has significantly advanced the availability of plant genome sequences, including that of the *Salacca* sp. Re-sequencing entire genomes, made possible by NGS technology, has increased knowledge of plant diversity. In contrast, traditional Sanger sequencing is costly (Visendi et al. 2014). As of May 30, 2024, the NCBI DNA Database contains 78 raw short read archive (SRA) data of 24 *Salacca* sp., with sizes ranging from 8.8 to 18,749.98 Mb nucleotides, and a complete Cp genome of *Salacca ramosiana* which can be used as a reference. Recently, whole genome sequences of *Salacca sumatrana* have been determined by (Matra et al. 2019) using shotgun sequencing. This research is a follow-up study using the generated *S. sumatrana* genome data and other *Salacca* species, focusing on the Cp genome diversity by assembling and examining the Cp genomes of seven *Salacca* species and determining their evolutionary relationships within the Arecaceae family. This paper aims to examine the Cp genomic data and explore the genetic variations of Cp genomes among *Salacca* species, which is essential for evolutionary studies. These analyses are critical for preserving genetic diversity and ensuring sustainable utilization of *Salacca* species in the future.

MATERIALS AND METHODS

Salacca genome SRA and Cp genome assembly

The raw SRA data for *S. sumatrana* were obtained from Dr. Deden Deradjat Matra, who sequenced and conducted a nuclear genome analysis of this species (Matra et al. 2019). Additionally, we searched the NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra/) for available raw SRA data and the NCBI Nucleotide DNA Database (https://www.ncbi.nlm.nih.gov/nucleotide/) for the available complete Cp genome of other *Salacca* species to perform an evolutionary classification based on the genomes.

The raw SRA data of the *Salacca* sp. genome were downloaded from NCBI database search results using the download reads from NCBI tool on the Galaxy software platform (https://usegalaxy.eu/). Data quality was assessed using the FASTQC tool integrated within the Galaxy platform (Jin et al. 2020). The downloaded raw SRA data of the *Salacca* genome contain a mixture of nuclear, chloroplastid, and mitochondrial genome sequences. Upon completing the download and quality control of the raw SRA genomic data, further analysis was done using the Get Organelle tool on the Galaxy platform to extract and assemble the Cp genome. The assembled Cp genomes were used in subsequent analysis. The nucleotide sequence analysis, organization, structure, and GC content of the identified *Salacca* Cp genomes, including the LSC, SSC, and IR regions, were determined using Geneious Prime 2019.1.1 version 11 (https://www.geneious.com). The gene counts were calculated with the assistance of a Microsoft Excel worksheet.

Chloroplast genome annotation

The previously assembled Cp genome of *Salacca* species was saved in the fasta file for genome annotation processes. The Cp genome annotation was conducted using CHLOROBOX (GeSeq-Annotation of Organeller Genomes) (https://chlorobox.mpimp-golm.mpg.de/geseq.html) through sequence similarity criteria of 95% for proteins, protein-coding DNA, rRNAs, and tRNAs. The tRNAscan-SE v2.0.3 with the default parameters was used to ensure accurate identification and annotation of the tRNA genes (Lowe and Chan 2016). The Organellar Genome DRAW (OGDRAW) tool was used to generate circular representations of the Cp genome maps for *S. sumatrana* and its closely related species, giving an informative Cp genome graphical representation (https://chlorobox.mpimpgolm.mpg.de/OGDraw.html).

Relative synonymous codon usage

The relative synonymous codon usage (RSCU) value was obtained by comparing the occurrence of a specific codon to the estimated frequency of the synonymous codon encoding similar amino acids. The frequency of codon usage is evaluated only for the regions of coding sequences (CDS) in all protein-encoding genes, utilizing the CAIcal tool, and the RSCU values were calculated manually using the CAIcal tool outputs (http://genomes.urv.es/CAIcal) (Rahmawati et al. 2021).

Comparative analysis of chloroplast genome

The base compositions and frequencies were determined using Geneious Prime - 2019.1.1 version 11 software (https://www.geneious.com) and estimated for the LSC, SSC, and IR sequences of seven *Salacca* species. The IRscope [(tool for visualizing the junction sites of the Cp genomes) (https://irscope.shinyapps.io/irapp/)] was used to analyze the inverted repeat (IR) changes (expansion or contraction) in the *Salacca* species Cp genomes by using the GeSeq outputs and imported GenBank accessions number from the NCBI (Amiryousefi et al. 2018).

SNP and insertion/deletion (InDels) quantification and spatial organization

Changes in nucleotide sequences occur primarily due to nucleotide substitution mutations resulting in single nucleotide polymorphisms (SNPs) (Deng et al. 2017) and insertion or deletion of DNA fragments, causing insertion-deletion (InDel) mutations (Sehn 2015). The frequency of the SNPs and the INDEL variants in the Cp genome may be used to develop genetic markers (Yang et al. 2016), which can differentiate accessions within and among *Salacca* species. The library of the determined SNPs and InDels may be essential in the genetic analysis of maternal inheritance, the intra- and inter-specific accessions differentiations, phylogeographic and phylogenetic analyses, gene-specific studies, and support for the future of *Salacca* breeding programs. A comprehensive examination of the number and distribution of SNPs and InDels within the Cp genomes is performed by evaluating the multiple-sequence alignment (MSA) outputs of the seven *Salacca* Cp genomes generated by the MAAFT alignment package (Yamada et al. 2016) in the Geneious Prime 2019.1.1 version 11 (https://www.geneious.com).

Quantity and distribution of SSRs

Using Phobos version 3.3.12(https://www.ruhr-uni-bochum.de/spezzoo/cm/cm_phobos_download.htm), the simple sequence repeat (SSR) comprising mono-, di-, tri-, and tetra-nucleotides are found inside the Cp genomes with the provided search parameters of a minimum of 8 repeat units for mononucleotide, a minimum of 4 repeat units for dinucleotide, and a minimum of 2 repeat units for trinucleotide repeats.

Phylogenetic analysis

Phylogenetic analysis using the Cp genome sequences was done for 45 members of the Arecaceae family, with *Dasypogon bromeliifolious* serving as an outgroup. All Cp genome multiple sequence alignment (MSA) were aligned using MAAFT (Yamada et al. 2016). Subsequently, the Tamura-Nei genetic distances and Neighbor-Joining tree construction methods were done to infer phylogenetics within the evaluated Arecaceae accessions. The bootstrap analysis with 1,000 replicates was used for statistical analysis. The MSA and phylogenetic analysis were conducted using Geneious Prime 2019.1.1 version 11 software (https://www.geneious.com).

RESULTS AND DISCUSSION

Salacca genome SRA

Seventy-eight accessions of the *Salacca* genome SRA were identified in the NCBI SRA database search results, comprising 24 *Salacca* species. The smallest SRA data is for *Salacca sarawakensis*, with a download data size of 8.8 Mb, and the number of bases is 70.1 Kb (Acc. No. ERX10665907). The two most extensive SRA data (Table 1) are for *S. zalacca* with a download data size of 18,750.0 Mb, and numbers of bases are 29.8 Gb (Acc. No. SRX14245468), and *S. sumatrana* with the download data size of 3,761.8 Mb, and numbers of bases of 5.9 Gb (Acc. No. DRX142533). The *Salacca* genome SRA with the download data size of 750.0 Mbytes and several bases of at least 2.0 Gbases were selected since they could assemble the complete Cp genome (Table 1).

Meanwhile, the other *Salacca* genome SRAs are less than 489.5 Mbytes in data size and 1.23 Gbases in base numbers (ERX10668233, *Salacca lophospatha*). Moreover, assembling the 489.5 Mbytes raw genomic SRA and 1.23 Gbases in base numbers resulted in recovering a fraction of the Cp genome. Therefore, the data size of less than 489.5 Mbytes and the base numbers of less than 1.23 Gbases were not included in the subsequence Cp genome assembly and analysis. The search for the Cp genome as a query for the nucleotide DNA database (https://www.ncbi.nlm.nih.gov/nuccore/) yielded only one accession of *S. ramosiana* complete Cp genome (Acc. No. KT312921.1). The complete Cp genome of *S. ramosiana* is 0.160 Mbytes in data size and 0.000149 Gbases in base numbers (Table 1). The Cp genome of *S. ramosiana* served as the reference for validating the assembly of Cp genomes in additional *Salacca* species.

Chloroplast genome assembly

The assembled Cp genome from raw SRA genome data of shotgun genome sequencing of *S. sumatrana* resulted in 157,936 bp nucleotide sequences. Meanwhile, the assembled Cp genome from raw data of the other six *Salacca* species shotgun genome sequencing ranges from the smallest, 157,047 bp (*S. ramosiana*), to the largest, 158,182 bp (*S. affinis*) (Table 2). Seven *Salacca* Cp genomes exhibit quadripartite structures comprising IR_A and IR_B, one LSC, and one SSC region. A representative of the whole Cp genome of *S. sumatrana* species is presented in Figure 1.

Table 1. *Salacca* species and their corresponding accession numbers, SRA data size, and remarks for the next generation sequences. The listed *Salacca* species represent the six largest sizes of *Salacca* genome SRA downloaded from the National Center for Biotechnology Information (NCBI)

Salacca species	SRA Acc. No.	SRA data		Remarks
		Download size (Mbytes)	Number of bases (Gbases)	
S. zalacca	SRX14245468	18,749.98	29.8	HiSeq X Ten paired-end sequencing; Raw reads: BKL055-skim.
S. sumatrana	DRX142533	3,761.76	5.9	BGI-Seq 500 paired-end sequencing of SAMD00136089
S. glabrescens	ERX10667619	362.87	3.62	HiSeq X Ten paired-end sequencing; Raw reads: BKL091-skim
*S. wallichiana**	SRX11969405	1,129.6	3.1	HiSeq X Ten paired-end sequencing; Raw reads: RBL117-skim
S. ramosiana	ERX10667693	1,017.68	2.8	Illumina TruSeq paired-end sequencing libraries: sequenced using the BGISEQ-500
S. secunda	ERX10667792	870.11	2.4	HiSeq X Ten paired-end sequencing; Raw reads: RBL117-skim
S. affinis	ERX10667685	782.39	2.1	HiSeq X Ten paired-end sequencing; Raw reads: BKL077-skim
S. ramosiana	KT312921.1	0.160	0.000149	*S. ramosiana*, complete plastid genome

Note: *Partial Cp genome

Table 2. Features of the chloroplast genome of seven *Salacca* species assembled from the downloaded whole-genome short-read archived (SRA) from the National Center for Biotechnology Information (NCBI) SRA database

Genome features	*S. affinis*	*S. sumatrana*	*S. glabrescens*	*S. zalacca*	*S. wallichiana**	*S. secunda*	*S. ramosiana*
Genome size (bp)	158,182	157,936	157,977	157,723	156,080	157,457	157,047
GC content %	37.2 %	37.2 %	37.3 %	37.3 %	37.3 %	37.4 %	37.4 %
LSC length(bp)	85,649	85,467	85,737	85,634	84,528	85,383	85,121
SSC length(bp)	17,861	17,751	17,862	17,723	17,420	17,725	17,594
IR length (bp)	27,336	27,359	27,189	27,183	27,002	27,178	27,166
GC content in LSC (%)	35.3 %	35.3 %	35.3 %	35.3 %	35.3 %	35.4 %	35.4 %
GC content in SSC (%)	31 %	30.8 %	31 %	31.2 %	31.2 %	31.2 %	31.3 %
GC content in IR (%)	42.2 %	42.2 %	42.4 %	42.4 %	42.5 %	42.5 %	42.4 %

Note: *Partial Cp genome

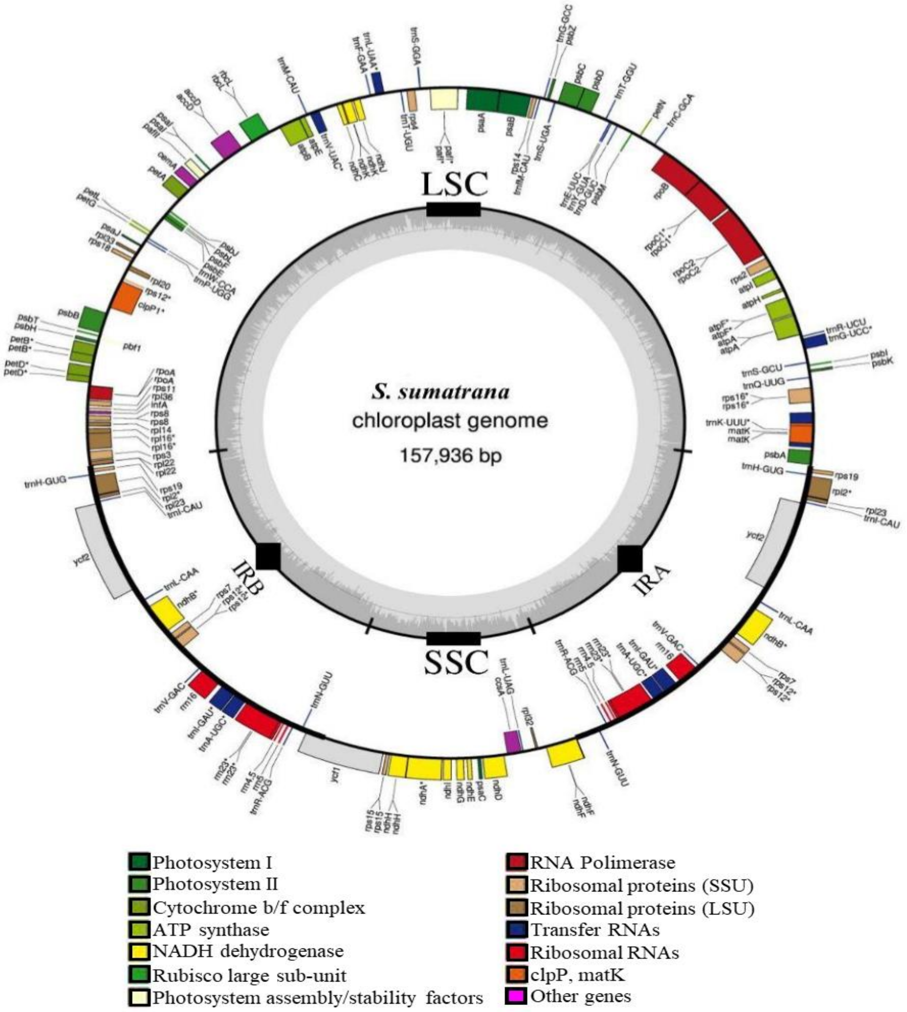

Figure 1. Circular diagram represents the Cp genome of *S. sumatrana*. Transcriptional activity occurs clockwise for genes outside the outer black circular line and counter-clockwise for genes inside. Genes for photosystems, cytochrome b/f complex, ATP synthase, NADH dehydrogenase, Rubisco, RNA polymerase, ribosomal proteins, transfer RNAs, ribosomal RNAs, genes for various functions, and conserved open reading frames (*ycf*) are among the functional groupings of genes that are represented by different colors.

The length of the LSC among the Cp genome of seven *Salacca* species varied from 85,737 bp in *S. affinis* to 85,121 bp in *S. ramosiana*. On the other hand, the length of SSC among the seven *Salacca* species varied from 17,862 bp (*S. glabrescens*) to 27,359 bp (*S. sumatrana*). In contrast, the IR varied from 17,594 bp (*S. ramosiana*) to 27,002 bp (*S. wallichiana*) (Table 2). The genetic composition and arrangement among the Cp genome of seven *Salacca* species exhibit a high degree of similarity, and they are aligned with the genetic architecture of the preserved flowering plant's Cp genomes (Wicke et al. 2011). The varying lengths of Cp genomes of *Apiales* are influenced by the constriction and expansion of IR region borders, as observed in angiosperms (Downie and Jansen 2015).

The IR regions have the highest GC concentration among the seven *Salacca* species studied, followed by the LSC and SSC regions (Table 2). Based on the assembled Cp genome of *S. sumatrana* species, the GC content of the IR region is 42.2 %, the LSC region is 35.3 %, and the SSC region is 30.8%. Among seven *Salacca* species, the IR region's GC content of *S. wallichiana* and *S. ramosiana* (42.5%) are the highest, while those of *S. affinis* and *S. sumatrana* are the lowest (42.2%). The LSC region's GC content of *S. secunda* and *S. ramosiana* is 35.4%, while for the other *Salacca* species is 35.3%. Meanwhile, the SSC region's GC content of *S. ramosiana* is the highest (31.3%), while *S. sumatrana* is the lowest (30.8%) (Table 2).

The Cp genomes' GC content varies among genes in different functional categories, with some genes having higher GC content than others (Green 2011). The highest to lowest GC content among functional genes in the Cp genome are ribosomal RNA genes, transfer RNA genes, photosynthetic genes, genetic system genes, and NADH-coding genes (Rahmawati et al. 2021). The lower GC contents of the LSC and SSC regions in all *Salacca* species than the IR region are attributed to ribosomal RNA (rRNA) in the IR region. A high GC content increases sequence complexity and contributes to the overall genome stability (Kaila et al. 2017). NADH dehydrogenase genes are the main reason for the lowest GC content in the SSC region, which has the lowest GC content than the LSC and the IRs (Jansen and Ruhlman 2012).

Chloroplast genome annotation

Four categories of genes are identified from annotating the assembled Cp genome of seven *Salacca* species, including self-replicating genes, genes for photosynthetic, genes for other functions, and genes of unknown function (Table S1). The annotation of the assembled Cp genome of *S. sumatrana* identified 140 functional gene copy numbers belonging to 24 gene groups and 112 gene ID numbers (Table 3). Meanwhile, the annotation of the assembled Cp genome of other *Salacca* species identified ranges from 79-140 functional gene copy numbers belonging to 24 gene groups and 73-113 gene ID numbers in the Cp genome (Table 4). Among the evaluated *Salacca* species, the number of protein-encoding genes ranges from 25 to 51, the tRNA gene ID from 48 - 62, and gene copy ranges from 51-71, and *S. wallichiana* has only tRNA 11 genes, and the same eight rRNA genes for all species except *S. wallichiana* has 4 (Table 3).

In the Cp genome of *S. sumatrana*, there are 23 duplicated genes, among the highest observed within the genus. Other *Salacca* species also show a range of duplicated genes in their Cp genomes: *S. affinis* (21), *S. glabrescens* (22), *S. zalacca* (21), *S. wallichiana* (4), *S. secunda* (8), and *S. ramosiana* (20) (Table 4). This variability in gene duplication across species highlights significant genetic diversity within the genus. Such differences may reflect varying evolutionary pressures and adaptations specific to each species' ecological niche. The higher number of duplicated genes in species like *S. sumatrana* and *S. glabrescens* could suggest an enhanced genomic robustness or a more complex genetic architecture, potentially conferring adaptive advantages. This genomic analysis provides a foundation for further exploration into the evolutionary dynamics and functional implications of gene duplication in *Salacca* species, with potential applications in conservation and agricultural optimization.

Genes within the chloroplast genome

Nine genes encoding the large ribosome sub-unit (*rpl2, rpl14, rpl16, rpl20, rpl22, rpl23, rpl32, rpl33,* and *rpl36*) are present in the Cp genome of *S. sumatrana* (Table 5). Most of the ribosomal genes are single copies. However, there are two copies of the *rpl2* and *rpl23* genes. Ribosomal genes in the Cp genome for the other *Salacca* species are the same as in *S. sumatrana*. However, a single copy of the *rpl2* and *rpl23* is present in the *S. wallichiana* and *S. secunda* Cp genome. Moreover, the *rpl20, rpl32,* and rpl33 genes are absent in the assembled Cp genome of *S. wallichiana* (Table 5).

The Cp genomes of *S. sumatrana* and *S. affinis* contain 29 genes encoding transfer RNAs. Moreover, *S. zalacca* and *S. secunda* Cp genomes have 30, while *S. glabrescent* and *S. ramosiana* Cp genomes have 27 transfer RNA genes. Some *trnA* genes occur in two or more copies in the Cp genome of *S. sumatrana*, such as *trnH-GUG, trnA-UGC, trnI-CAU, trnI-GAU, trnL-CAA, trnN-GUU, trnR-ACG* and *trnV-GAC* (Table 4). The same *trn* genes are also present in two or more copies in *S. affinis*, *S. glabrescens*, *S. zalacca* and *S. ramosiana*. On the other hand, those transfer RNA genes exist as single copies in *S. wallichiana* and *S. secunda*. Only eleven transfer RNA genes are identified from the partial assembly of the *S. wallichiana* Cp genome (Table 4). Twenty of the tRNA genes are absent from the *S. wallichiana* Cp genome due to the incomplete assembly.

Table 3. Gene counts in the chloroplast genomes of various *Salacca* species

Salacca species	Protein-coding genes		RNA-coding genes		Total	
	Gene ID no.	Gene copy no.	Gene ID no.	Gene copy no.	Gene ID no.	Gene copy no.
S. affinis	50	65	62	71	112	136
S. sumatrana	50	69	62	71	112	140
S. glabrescens	47	65	61	70	108	135
S. zalacca	51	64	62	71	113	135
*wallichiana**	25	28	48	51	73	79
S. secunda	51	52	61	68	112	120
S. ramosiana	47	63	62	69	109	132

Note: *****Partial Cp genome

Table 4. Gene duplication in subunit gene IDs across various *Salacca* species/functional genes with two or more physical copies in the Cp genome of seven *Salacca* species

Gene categories and	Number of duplicate genes in *Salacca* species (copy numbers)						
gene ID	*S. affinis*	*S. sumatrana*	*S. glabrescens*	*S. zalacca*	*S. wallichiana**	*S. secunda*	*S. ramosiana*
Large sub-unit of ribosome							
rpl2	2	2	2	2	1	1	2
rpl23	2	2	2	2	1	1	2
Transfer RNA genes							
trnA-UGC	2	2	2	2	1	1	2
trnG-UCC	1	1	1	1		1	2
trnH-GUG	2	2	2	2	1	1	2
trnl-CAU		2		2	1	1	
trnI-GAU	4	4	4	2	1	1	2
trnL-CAA	2	2	2	1	1	1	1
trnM-CAU	1	4	4	1	1	1	3
trnN-GUU	2	2	2	2	1	1	2
trnR-ACG	2	2	2	2	1	1	2
trnV-GAC	2	2	2	2	1	1	2
Ribosomal RNA genes							
rrn16s	2	2	2	2	1	2	2
rrn23s	2	2	2	2	1	2	2
rrn4.5s	2	2	2	2	1	2	2
rrn5s	2	2	2	2	1	2	2
Small subunits of ribosomes							
rps12	3	3	3	3	2	2	3
rps19	2	2	2	2	1	1	2
rps7	2	2	2	2	1	1	2
rps8	1	1	1	1	2	1	1
Subunit of NADH-Dehydrogenase							
ndhB	2	2	2	2	1	1	2
ndhK	2	2	2	2	1	2	1
Subunit of Cytochrome b/f complex							
petD	2	2	2	2	1	2	1
Conserved open reading frames							
ycf1	2	2	2	2	2	2	2
ycf2	2	2	2	2	2	1	2

Note: *Partial Cp genome

Six *Salacca* species have two copies of the ribosomal RNA genes (*rrn23S, 16S, 5S,* and *4.5S*), while *S. wallichiana* has only one copy of the four ribosomal genes. Twelve genes encoding small subunits of ribosome (*rps*) are found in the assembled Cp genome of *S. sumatrana, S. affinis, S. zalacca,* and *S. secunda* (Table 5). One of the twelve rps genes is missing from the assembled Cp genome of *S. glabrescens* (*rps16*) and *S. ramosiana* (*rps4*). Meanwhile, four *rps* genes (*rps11, rps14, rps16,* and *rps18*) are missing from the assembled Cp genome of *S. wallichiana*. Two *rps* genes (*rps19* and *rps 7*) are present in two copies in Cp genome of *S. affinis, S. sumatrana, S. glabrescens, S. zalacca,* and *S. ramosiana*, and in single copy in *S. wallichiana* and *S. secunda*. The *rps12* is present in three copies in *S. affinis, S. sumatrana, S. glabrescens, S. zalacca,* and *S. ramosiana*, and in two copies in *S. wallichiana* and *S. secunda* (Table 5). Seven *Salacca* species have a single copy of the four genes encoding sub-units of DNA-dependent RNA polymerase (*rpoA, B, C1,* and *C2*).

Eleven genes encoding subunits of NADH-dehydrogenase (*ndh*) are found in the assembled Cp genome of *S. sumatrana, S. affinis, S. glabrescens, S. zalacca, S. secunda,* and *S. ramosiana*. Two copies of the *ndhB* are present in *S. sumatrana, S. affinis, S. glabrescens, S. zalacca,* and *S. ramosiana* and one copy is present in *S. wallichiana* and *S. secunda* (Table 5). Meanwhile, two copies of the *ndhK* is present in *S. sumatrana, S. affinis, S. glabrescens, S. zalacca,* and *S. secunda,*and one copy is present in *S. wallichiana* and *S. ramosiana.* The *ndhE* and *ndhG* are not identified in the assembled CP genome of *S. wallichiana.* Five single-copy genes encoding the subunits of photosystem I (*psaA, B, C, I,* and *J* are present in the Cp genome of six *Salacca* species (Table 5). However, *psaJ* is missing in the assembled Cp genome of *S. wallichiana.*

Fourteen single-copy genes encoding the subunits of photosystem II (*psaA, B, C, D, E, F, G, H, I, J, K, L, M, T,* and *Z*) are present in the Cp genome of *S. sumatrana, S. affinis, S. zalacca,* and *S secunda.* In contrast, 15 single-copy *psb* genes are present in *S. ramosiana* (Table 5). The *psbN* is only present in the assembled Cp genome of *S. ramosiana* and absent in the assembled Cp genome of other *Salacca* species. Moreover, *psbA* is missing in the assembled Cp genome of *S. glabrescens,* and six *psb* genes (*psbA, F, J, L, M,* and *N*) are missing in the assembled Cp genome of *S. wallichiana* (Table 5). There are two copies of *psbC* in *S. wallichiana* but only one in other species.

Six genes encoding the subunits of cytochrome b/f (*petA, B, D, G, L,* and *N*) are present in the assembled Cp genome of *S. sumatrana, S. affinis, S. glabrescens, S. zalacca, S. secunda,* and *S ramosiana*. Two copies of the *petD* is present in the assembled Cp genome of *S. sumatrana, S. affinis, S. glabrescens, S. zalacca,* and *S. secunda,* and single copy of *petD* is present in assembled Cp genome of *S. wallichiana* and *S. ramosiana*. Moreover, *petB, petG,* and *petL* are missing in the assembled Cp genome of *S. wallichiana* and *S. ramosiana* (Table 5). Six single-copy genes encoding the subunit of ATP synthase (*atpA, B, E, F, H,* and *I*) are present in the assembled Cp genome of *S. sumatrana, S. affinis, S. glabrescens, S. zalacca, S. secunda,* and *S. ramosiana*. However, *atpI* is missing in the assembled Cp genome of *S. wallichiana* (Table 5).

Single copies of genes encoding large subunits of rubisco (*rbcL*), maturase (*matK*), and envelope membrane protein (*cemA*) are present in the assembled Cp genome of seven *Salacca* species. Similarly, a single copy of genes encoding a subunit of acetyl-CoA carboxylase (*accD*), C-type cytochrome synthesis gene (*ccsA*), and translation initiation factor (*infA*) are present in *S. sumatrana, S. affinis, S. glabrescens, S. zalacca, S. secunda* and *S. ramosiana* (Table 5). In the assembled Cp genome of *S. wallichiana*, accD and *infA* are present, while *ccsA* is absent. Seven genes (i.e., conserved open reading frame) of unknown functions were identified in the assembled CP genome of the *Salacca* species. Two copies of the conserved *ycf1* and *ycf2* open reading frames are present in the assembled Cp genome of six *Salacca* species (Table 5). Two copies of *ycf1* and a single copy of ycf2 were found in the assembled Cp genome of *S. secunda*. The *ycf3* and *ycf4* open reading frames are only present in the assembled Cp genome of *S. ramosiana* and absent in other *Salacca* species. Single copies of *pafl* and *pafl1* open reading frames are found in the assembled Cp genome of six *Salacca* species and are lacking in *S. ramosiana*.

Table S1. Genes encoded by the *Salacca sumatrana* Cp genome

Category of genes	**Group of gene**					
Self-replication	Ribosomal RNA genes	*rrn23s(x2)*	*rrn16s(x2)*	*rrn4.5s(x2)*	*rrn5s(x2)*	

	Transfer RNA genes	*trnV-UAC,*
		trnS-GGA,
		trnG-GCC,
		trnL-UAA,
		trnF-GAA,
		trnL-UAG,
		trnW-ssCCA,
		trnP-UGG,
		trnL-CAA,
		trnl-GAU
			trnM-CAU,
			trnT-UGU,
			trnC-GCA,
			trnI-GAU,
			trnY-GUA,
			trnD-GUC,
			trnR-UCU,
			trnG-UCC,
			trnS-GCU,
			trnQ-UUG
				trnfM-CAU,
				trnT-GGU,
				trnS-UGA,
				trnE-UUC,
				trnK-UUU
					trnA-UGC(x2)
					trnV-GAC(x2)
					trnN-GUU(x2)
					trnR-ACG(x2)
					trnH-GUG(x2),
					trnl-CAU(x2)	

	Small Subunits of ribosomes	*rps2* *rps11* *rps18*	*rps3* *rps12 (x3)* *rps19(x2)*	*rps4* *rps14*	*rps7(x2) rps15*	*rps8* *rps1*
	Large Subunit of ribosomes	*rpl 2(x2)* *rpl 23(x2)*	*rpl14* *rpl32*	*Rpl16* *rpl33*	*rpl20* *rpl36*	*rpl22*
	DNA-dependent RNA polymerase	*rpoA*	*rpoB*	*rpoC1*	*rpoC2*	
	Subunit of NADH-Dehydrogenase	*ndhA* *ndhF* *ndhK(x2)*	*ndhB(x2)* *ndhG*	*ndhC, ndhH*	*ndh D,ndhI*	*ndhE,* *ndhJ*
	Subunit of photosystem I	*psaA*	*psaB*	*psaJ*	*psaI*	*psaC*
	Subunit of photosystem II	*psbA* *psbF* *psbL*	*psbB* *psbH* *psbM*	*psbC* *psb I*	*psbD* *psbJ* *psbT*	*psbE* *psbK* *psbZ*
	Subunit of Cytochrome b/f complex	*petA*	*petB*	*petD(x2)*	*petL,*	*petN* *petG*
Genes for photosynthesis	Subunit of ATP synthase	*atpA* *atpI*	*atpB*	*atpE*	*atpF*	*atpH*
	Subunits of rubisco	*Rbcl*				
	Maturase	*matK*				
	Envelope membrane protein	*cemA*				
Others	Subunit of acetyl-CoA Carboxylase	*accD*				
	C-type Cytochrome synthesis gene	*ccsA*				
	Translational initiation factor	*infA*				
Genes of unknown function	Conserved open reading frames	*ycf2(x2)*	*ycf1*	*pafl* *pafll*	*pbf1*	
	Protease	*clpP1*				

Table 5. Category and group of genes with gene products and copy numbers across the *Salacca* species

Category and group of genes		**Gene products**	***S. affinis***		***S. sumatrana***		***S. glabrescens***		***S. zalacca***		***S. wallichiana****		***S. secunda***		***S. ramosiana***	
			SID	**CN**	**SID**	**CN**	**SID**	**CN**	**SID**	**CN**	**SID**	**CN**	**SID**	**CN**	**SID**	**CN**
Self-replicating DNA																
1	Large sub-unit of a ribosome	RNA	9	11	9	11	9	11	9	11	6	6	9	9	9	11
2	Transfer RNA genes	RNA	29	38	29	42	27	39	30	37	11	11	30	30	27	37
3	Ribosomal RNA genes	Protein	4	8	4	8	4	8	4	8	4	4	4	8	4	8
4	Small Subunits of ribosomes	RNA	12	16	12	16	11	15	12	16	8	11	12	13	11	15
5	DNA-dependent RNA polymerase	Protein	4	4	4	4	4	4	4	4	4	4	4	4	4	4
6	Subunit of NADH-Dehydrogenase	Protein	11	13	11	13	11	13	11	13	9	9	11	12	11	12
7	Subunit of photosystem I	Protein	5	5	5	5	5	5	5	5	4	4	5	5	5	5
8	Subunit of photosystem II	Protein	14	14	14	14	13	13	14	14	9	10	14	14	15	15
9	Subunit of Cytochrome b/f complex	Protein	6	7	6	7	6	7	6	7	3	3	6	7	6	6
Genes for photosynthesis																
1	Subunit of ATP synthase	Protein	6	6	6	6	6	6	6	6	5	5	6	6	6	6
2	Envelope membrane protein	Protein	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	Maturase	Protein	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	Subunits of rubisco	Protein	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Genes for other function than photosynthesis																
1	Subunit of acetyl-CoA Carboxylase	Protein	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	C-type Cytochrome synthesis gene	Protein	1	1	1	1	1	1	1	1	NA	NA	1	1	1	1
3	Translational initiation Factor	Protein	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Genes of unknown function																
1	Conserved ORF															
	pafl	Protein	1	1	1	1	1	1	1	1	1	1	1	1	NA	NA
	pafll	Protein	1	1	1	1	1	1	1	1	1	1	1	1	NA	NA
	pbf1	Protein	1	1	1	1	1	1	1	1	NA	NA	NA	NA	NA	NA
	ycf1	Protein	1	2	1	2	1	2	1	2	1	2	1	2	1	2
	ycf2	Protein	1	2	1	2	1	2	1	2	1	2	1	1	1	2
	ycf3	Protein	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	1
	ycf4	Protein	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	1
2	Protease	Protein	1	1	1	1	1	1	1	1	1	1	1	1	1	1
TOTAL			112	136	112	140	108	135	113	135	73	79	112	120	109	132

Note: SID: Subunit gene ID, CN: Copy Number, *****Partial Cp genome

Moreover, the *pbf1* open reading frame is present in the assembled Cp genome of four *Salacca* species and missing in the *S. wallichiana*, *S. secunda*, and *S. ramosiana* (Table 5). The *Salacca* Cp genomes exhibit minor variations, particularly in the gene count, despite the generally acknowledged significant conservation of Cp genomes in terrestrial plants (Kumar et al. 2016). The presence or absence of a particular gene among various *Salacca* species may be attributed to its translocation to the nucleus (Huang et al. 2017).

The *rps12* gene has been determined to undergo trans-splicing. The *rps12* gene undergoes trans-splicing due to the presence of one of its exons in the LSC area and the 3'-tail in the IR region. This phenomenon is also seen in *P. dactylifera* (Yang et al. 2010). The presence of the *ycf1* pseudogene has been documented in both *Byrsonima coccolobifolia* and *B. crassifolia* species. The *ycf1* gene, which covers 5745 base pairs, originated from the SSC region and migrated towards the SSC/IR$_A$ barrier. The *ycf1* gene duplication occurs at the 3'-end in the IR$_B$ region, forming a *ycf1* pseudogene that is 1389 base pairs long (Menezes et al. 2018). Like *S. sumatrana*, the 5559 bp *ycf1* gene began from the SSC area and then migrated towards the SSC / IR$_A$ boundary, forming a *ycf1* pseudogene (1346 bp).

A total of eight genes are found to include introns. The *clp*P1 gene has two introns among these genes; the seven remaining genes (*atp*F, *ndh*A, *pet*D, *rpo*C1, *rpl*16, *rps*16, and *paf*1) each have a single intron. Five tRNA genes have an tintron *trn*K-UUU, *trn*S-CGA, *trn*E-UUL, *trn*L-UAA, and *trn*A-UGC. The gene *trn*K-UUU contains the most extensive intron, measuring 2,586 base pairs of all the genes. The maturase (*mat*K) gene is also present in a *trn*K-UUU intron with a 1,554 bp. The tRNA gene *trn*K-UUU stands out from other tRNA genes due to its unique inclusion of the *mat*K gene within its intron. The *mat*K gene plays a crucial function in splicing the *trn*K intron in which it is present. The *mat*K gene is known to have a significant degree of divergence (Wilson 2004). Thus, it can be used as an indicator for examining evolutionary relationships within and between species (Harnelly et al. 2018).

Codon usage analysis

An RSCU score above 1.00 is regarded as high-frequency, and a score below 1.00 is considered less frequent in the genome (Gun et al. 2018). The genes encoding protein in *S. zalacca, S. affinis, S. sumatrana, S. ramosiana, S. glabrescens, S. wallichiana* and *S. secunda* consist of 47890, 46466, 45142, 46190, 46220, 22573, and 35999 codons, respectively. The RSCU calculation results among the *Salacca* species indicate the scores of 21, 16, 21, 24, 24, 23, 22, and 22, categorized as high frequency (Table 6). Several amino acids, such as Phenylalanine (F), VaIine (V), Serine (S), Tyrosine (Y), Cysteine (C), and Glycine (G), are absent in *S. affinis*. The codons ATG and TGG, which encode methionine (M) and tryptophan (T), are assigned an RSCU value of 1.00 because they represent both amino acids. A bias influences another amino acid because many synonymous codons can encode them, except Methionine (M) and Tryptophan (T). Leucine (L) is the most often utilized codon in the *Salacca* Cp genome, with its highest frequency in *S. glabrescens* at 5368 and *S. ramosiana* at 5280. Conversely, cysteine (C) is the least commonly used codon in the *Salacca* Cp genome, with 905 occurrences in *S. secunda* and 1140 in *S. ramosiana* (Table 7).

The RSCU values for the synonymous leucine codons are as follows: TTA = 1.404, TTG = 1.271, CTT = 1.285, CTC = 0.737, CTA = 0.98, and CTG = 0.624 (Table 7). The analysis reveals that the Cp genome exhibits a higher frequency of utilizing the TTA codon for leucine translation than other synonymous codons. The RSCU value recorded for TTA (1.404) is the highest. The codon CTG has the lowest frequency of usage, as indicated by its RSCU value of 0.624 in *S. zalacca*. The common initiator codon ATG is for most protein-coding genes in *Salacca* species (Table 7). However, there are several alternative start codons present in *Salacca* protein-coding genes. Specifically, eight genes utilize alternate start codons. For example, *cem*A and *pet*B use ATG, *rpl*16 and *rps*16 use CTA and TTA, respectively; *ndh*D uses CTA, *rps*12 uses ACT, *rpl*2 uses ACG, and *rps*19 uses GTG. Alternative codons act as start codons inside protein-coding genes (Wang et al. 2016). ACG and GTG are the start codons for *rpl*2 and *rps*19, respectively, in *Oryza sativa* (Liu et al. 2020).

Base composition and frequency count

The base/nucleotide composition analysis of LSC regions of the seven Cp genomes showed that *S. zalacca, S. ramosiana,* and *S. secunda* have the same proportion of A (31.80%) and G (17.30%). On the other hand, the lowest proportion of T and C nucleotides are 32.80% and 18.10%. In the context of SSC, sinensis exhibits the highest proportion of A (34.60%) nucleotide, whereas *S. sumatrana* and *S. glabrescens* exhibit the highest proportion of T nucleotides (34.70%) and C nucleotides (16.20%), for *S. ramosiana.* The overall adenine-thymine (A+T) composition exceeds 50% in comparison to the guanine-cytosine (G+C) content (Table 8). This study demonstrates that *Salacca* Cp genomes exhibit elevated amounts of A+T content, a characteristic commonly reported in the Cp genomes of angiosperm species (Bi et al. 2018). After analyzing the frequency of nucleotides in the seven Cp genomes, it shows that in the LSC region, the A and T nucleotides for all seven species are high compared to the SSC and IR regions. At the same time, the lowest counts, A, T, G, and C, are calculated in the SSC region (Table 9).

SNP and insertion/deletion (InDels) quantification and spatial organization

In the seven *Salacca* cp genomes, 813 SNPs have been found. The non-coding regions have 498 SNPs, 112 synonymous SNPs, 203 non-synonymous SNPs, and indels are 369. The frequency of non-synonymous SNPs is greater than synonymous SNPs in the percentage of SNPs and InDels (Figure 2). The analysis of SNP distribution reveals the presence of both non-synonymous and synonymous SNPs in nearly all protein-coding genes.

Table 6. Relative synonymous codon usage (RSCU) recognition pattern of all *Salacca* chloroplast genome

Amino acids	Codons	RSCU score in Cp genome of *Salacca* species						
		S. zalacca	*S. affinis*	*S. sumatrana*	*S. ramosiana*	*S. glabrescens*	*S. wallichiana**	*S. secunda*
Phe/F	TTT	2.66	-	2.74	2.71	2.54	2.30	2.52
	TTC	1.78	-	1.73	1.69	1.74	1.55	1.69
Leu/L	TTA	1.40	1.31	1.37	1.39	1.41	1.33	1.45
	TTG	1.28	1.32	1.22	1.37	1.33	1.06	1.32
	CTT	1.29	0.79	1.28	1.36	1.27	1.17	1.13
	CTC	0.73	0.81	0.85	0.89	0.81	0.81	0.70
	CTA	0.98	0.93	0.94	1.01	1.05	0.90	0.87
	CTG	0.62	0.63	0.56	0.64	0.65	0.62	0.57
Ile/I	ATT	2.22	2.19	2.33	2.22	2.15	1.86	2.15
	ATC	1.35	1.42	1.44	1.51	1.44	1.30	1.41
	ATA	1.83	1.96	1.84	2.03	1.87	1.46	2.02
VaI/V	GTT	1.00	-	0.99	1.06	0.98	1.06	0.97
	GTC	0.60	-	0.53	0.55	0.51	0.52	0.46
	GTA	0.87	0.93	0.86	0.96	0.89	0.95	0.95
	GTG	0.48	0.53	0.44	0.57	0.53	0.66	0.52
Ser/S	TCT	1.51	-	1.39	1.42	1.40	1.47	1.35
	TCC	1.11	-	1.10	1.16	1.07	1.23	1.03
	TCA	1.26	1.22	1.20	1.22	1.16	1.24	1.26
	TCG	0.81	0.75	0.75	0.78	0.70	0.76	0.72
	AGT	0.9	0.78	0.84	0.85	0.80	0.85	0.78
	AGC	0.57	0.55	0.61	0.58	0.58	0.74	0.56
Pro/P	CCT	0.78	0.79	0.81	0.86	0.75	0.87	0.74
	CCC	0.71	0.73	0.72	0.76	0.69	0.72	0.67
	CCA	0.91	1.00	0.90	0.98	0.88	0.97	0.89
	CCG	0.50	0.49	0.46	0.514	0.50	0.61	0.44
Thr/T	ACT	0.90	0.81	0.92	0.85	0.81	0.71	0.88
	ACC	0.74	0.69	0.71	0.72	0.66	0.80	0.67
	ACA	0.9	0.85	0.89	0.92	0.85	0.93	0.96
	ACG	0.45	0.49	0.49	0.49	0.44	0.59	0.47
Ala/A	GCT	0.60	-	0.61	0.62	0.54	0.72	0.62
	GCC	0.42	-	0.41	0.43	0.39	0.50	0.45
	GCA	0.60	0.56	0.52	0.55	0.49	0.67	0.60
	GCG	0.27	0.27	0.28	0.29	0.25	0.40	0.25
Tyr/Y	TAT	1.96	-	2.04	1.91	1.96	1.52	1.89
	TAC	0.92	-	0.95	0.91	1.00	0.91	0.89
His/H	CAT	1.16	1.21	1.17	1.21	1.22	1.07	1.05
	CAC	0.47	0.46	0.49	0.53	0.55	0.60	0.44
Gln/Q	CAA	1.27	1.24	1.34	1.31	1.34	1.26	1.31
	CAG	0.62	0.61	0.60	0.61	0.63	0.72	0.54
Asn/N	AAT	2.24	2.21	2.20	2.19	2.09	1.87	2.19
	AAC	0.95	1.10	0.98	1.02	1.01	0.94	1.00
Lys/K	AAA	2.62	2.58	2.53	2.60	2.50	2.17	2.79
	AAG	1.23	1.25	1.18	1.23	1.16	1.29	1.26
Asp/D	GAT	1.40	-	1.43	1.42	1.36	1.46	1.31
	GAC	0.54	0.54	0.54	0.52	0.50	0.62	0.52
Glu/E	GAA	1.80	1.76	1.81	1.17	1.61	1.72	1.91
	GAG	0.69	0.74	0.71	0.77	0.75	1.00	0.70
Cys/C	TGT	0.89	-	0.95	0.91	0.91	0.87	0.81
	TGC	0.56	-	0.57	0.53	0.53	0.60	0.54
Arg/R	CGT	0.51	0.46	0.51	0.47	0.46	0.57	0.49
	CGC	0.28	0.30	0.28	0.29	0.31	0.34	0.25
	CGA	0.79	-	0.74	0.70	0.72	0.85	0.72
	CGG	0.47	0.46	0.46	0.47	0.50	0.61	0.41
	AGA	1.39	1.32	1.38	1.31	1.36	1.64	1.50
	AGG	0.72	0.75	0.80	0.77	0.76	0.93	0.68
Gly/G	GGT	0.74	-	0.71	0.72	0.68	0.80	0.717
	GGC	0.40	-	0.43	0.39	0.39	0.60	0.39
	GGA	1.12	1.03	1.11	1.07	1.02	1.25	1.16
	GGG	0.64	0.65	0.65	0.65	0.67	0.87	0.63

Note: Bold means the RSCU value is more significant than 1.00, *****Partial Cp genome

Table 7. Number of codon usage recognition patterns in *Salacca* chloroplast genome

Amino acid	Codon	Number of codon usage in *Salacca* species						
		S. zalacca	*S. affinis*	*S. sumatrana*	*S. ramosiana*	*S. glabrescens*	*S. wallichiana**	*S. secunda*
Phe/F	TTT	2148	2160	2,216	2,153	2,091	896	1,697
	TTC	1450	1537	1,400	1,337	1,428	604	1,139
Leu/L	TTA	1136	1065	1,106	1,104	1,158	519	976
	TTG	1028	1074	991	1,090	1,098	414	885
	CTT	1039	1037	1,038	1,075	1,042	455	760
	CTC	596	659	690	703	670	316	471
	CTA	793	756	760	802	863	352	584
	CTG	505	508	455	506	537	243	384
Ile/I	ATT	1800	1780	1,887	1,761	1,765	727	1,443
	ATC	1094	1152	1,169	1,200	1,187	508	945
	ATA	1477	1587	1,492	1,610	1,540	569	1,362
VaI/V	GTT	807	821	803	813	803	412	653
	GTC	445	428	431	439	422	201	312
	GTA	697	753	698	765	734	370	639
	GTG	389	428	362	452	438	259	346
Ser/S	TCT	1223	1152	1,123	1,123	1,153	575	910
	TCC	899	935	893	918	884	481	696
	TCA	1025	994	970	966	954	483	847
	TCG	656	612	608	620	573	296	487
	AGT	706	634	681	671	658	332	536
	AGC	459	447	496	461	479	290	374
Pro/P	CCT	632	641	653	680	620	341	494
	CCC	581	591	580	599	570	281	447
	CCA	741	817	727	777	723	379	588
	CCG	404	401	372	408	411	238	295
Thr/T	ACT	725	653	748	673	665	278	594
	ACC	603	561	577	574	542	311	451
	ACA	699	692	717	732	698	366	644
	ACG	363	394	398	389	361	229	315
Ala/A	GCT	483	452	492	494	441	283	416
	GCC	346	339	336	340	317	193	304
	GCA	470	454	422	434	402	260	397
	GCG	218	217	230	230	209	155	168
Tyr/Y	TAT	1588	1602	1,649	1,514	1,614	593	1,270
	TAC	749	760	767	724	820	357	599
His/H	CAT	939	985	945	960	1,007	418	706
	CAC	383	376	397	417	451	232	293
Gln/Q	CAA	1026	1009	1,083	1,036	1,101	493	883
	CAG	505	494	487	485	522	279	366
Asn/N	AAT	1817	1790	1,778	1,737	1,723	730	1,472
	AAC	769	827	791	810	835	367	676
Lys/K	AAA	2121	2094	2,050	2,065	2,049	845	1,873
	AAG	999	1011	954	978	952	504	849
Asp/D	GAT	1128	1132	1,161	1,126	1,120	568	884
	GAC	443	440	436	412	410	241	348
Glu/E	GAA	1462	1425	1,465	1,358	1,326	671	1,283
	GAG	557	598	575	608	618	389	467
Cys/C	TGT	723	782	769	721	750	338	541
	TGC	450	468	463	419	436	234	364
Arg/R	CGT	416	372	417	376	382	221	326
	CGC	230	244	229	229	258	131	167
	CGA	637	611	597	557	593	330	486
	CGG	378	376	372	376	415	237	276
	AGA	1122	1073	1,114	1,037	1,122	641	1,009
	AGG	582	604	650	610	628	361	456
Gly/G	GGT	595	584	578	576	560	311	482
	GGC	322	324	348	306	317	235	259
	GGA	907	833	897	849	840	488	779
	GGG	519	523	530	519	549	340	420

Note: *****Partial Cp genome

Table 8. Quantitative examination of the proportion of nucleotides in various structural parts of the cp genomes of seven *Salacca* species

Species	Percentage (%) of bases in the structural parts of the Cp genome											
	LSC				SSC				IR			
	A	T	G	C	A	T	G	C	A	T	G	C
S. affinis	32	33	17	18	34	35	15	16	29	29	20	22
S. sumatrana	32	33	17	18	35	35	15	16	29	29	20	22
S. glabrescens	32	33	17	18	35	35	15	16	29	29	20	22
S. zalacca	32	33	17	18	35	34	16	15	29	29	21	22
*S. wallichiana**	32	33	17	18	34	35	15	16	29	29	21	22
S. secunda	32	33	17	18	34	35	15	16	29	29	21	22
S. ramosiana	32	33	17	18	34	35	15	16	29	29	21	22

Note: ***Partial Cp genome

Table 9. Quantitative examination of the frequency of nucleotides in various structural parts of the cp genomes of seven *Salacca* species

Species	Frequency of bases in the structural parts of the Cp genome											
	LSC				SSC				IR			
	A	T	G	C	A	T	G	C	A	T	G	C
S. affinis	27274	28150	14761	15464	6152	6179	2667	2863	7849	7944	5580	5963
S. sumatrana	27186	28081	14742	15458	6124	6153	2643	2831	7856	7953	5590	5960
S. glabrescens	27186	28081	14742	15458	6124	6153	2643	2831	7856	7953	5590	5960
S. zalacca	27262	28103	14777	15492	6126	6063	2864	2670	7775	7878	5577	5953
*S. wallichiana**	26968	27757	14609	15295	5954	6025	2628	2804	7691	7804	5542	5922
S. secunda	27036	27928	14746	15440	6054	6114	2667	2855	7770	7865	5583	5960
S. ramosiana	27082	27891	14699	15447	6015	6068	2655	2856	7772	7866	5573	5955

Note: ***Partial Cp genome

Table 10. Non-synonymous and synonymous SNPs in each gene of the *Salacca* Cp genome

Genes	Synonymous SNPs	Non-synonymous SNPs	Genes	Synonymous SNPs	Non-synonymous SNPs	Genes	Synonymous SNPs	Non-synonymous SNPs
accD	2	4	*petB*	3	2	*rpoA*	1	3
atpA	2	3	*petD*	2	1	*rpoB*	3	6
atpH	0	1	*psbK*	0	1	*rpoC2*	9	10
atpE	1	0	*psbM*	0	2	*rps3*	1	4
atpI	0	2	*psbD*	1	2	*rps4*	0	1
atpB	2	4	*psbC*	2	1	*rps8*	0	1
cemA	3	2	*psbZ*	0	1	*rps11*	0	2
ccsA	2	1	*psbJ*	0	2	*rps12*	0	1
clp1	0	3	*psbL*	0	1	*rps14*	0	1
infA	0	1	*psbB*	6	1	*rps16*	0	3
matk	2	10	*psbT*	1	0	*rps19*	2	2
ndhF	7	15	*psbH*	2	0	*psaA*	1	8
ndhK	2	1	*rbcL*	8	6	*psaI*	1	0
ndhD	2	3	*rpl2*	1	0	*pafl*	1	2
ndhG	1	1	*rpl14*	0	1	*pafll*	0	1
ndhE	1	2	*rpl16*	0	1	*ycf1*	21	38
ndhI	2	2	*rpl20*	0	1	*ycf2*	9	20
ndhA	1	6	*rpl22*	2	1	*ycf4*	1	0
ndhH	1	2	*rpl33*	0	1			

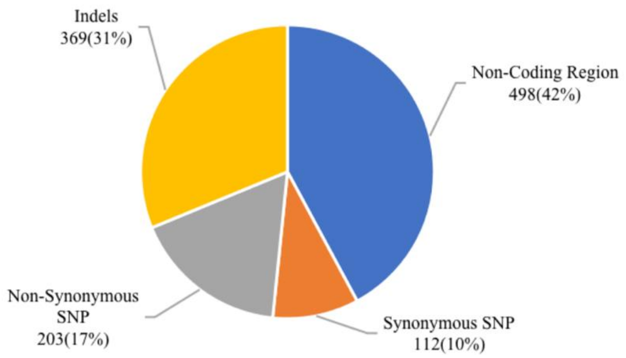

Figure 2. Percentage of SNPs and indels in Cp genomes of *Salacca*

In the following genes, only non-synonymous SNPs are found *rps16, psbK, atpH,atpL, psbM, psbZ, rps14, rps4, pafll, psbJ, psbJ, rpl33, rpl20, rps12, clp1, rps11, rps8, infA, rpl14, rpl16* (Table 10). No SNP is detected in the following genes: *petG, petN, petL, petA, ndhC, ndhJ, psbA, psbF, psbI, psbE,* and *rpl36*. Some SNPs are found in introns of the following genes: *atpF, rpoC1, pafl i*ntron 2, *clp1* intron 2, *petD* intron 1, and *rpl16* intron 1. Some SNPs are found in the exon regions like *rpoC2* in exon 2 (3 synonymous, six non-synonymous), *pafl* in exon 2 (1 synonymous, two non-synonymous), *rps12* in exon 1 one non-synonymous, *clp1* exon 2 one non-synonymous, *petB* in exon 2 has two synonymous SNPs, *rpl16* in exon two has one synonymous SNP, *ndhA* in exon two has one synonymous and two non-synonymous and in exon 2 has five non-synonymous SNPs and *rpl2* in exon 2 has one synonymous SNP.

A total of 369 InDels are identified in all *Salacca* Cp genomes. InDels in protein-coding regions are the following, with genes having InDels *psbA, psbB, psbE, psbF, accD, atpB, atpH, ccsA, matk, ndhD, petB, petD, rpoA, rpoC1, infA, ycf1, ycf2, rps15, rps12, rpl16, rpl22*. InDels are also detected in atpF, clpP1, ndhA, and petB introns. The genes with no InDels found are following *rbcL, atpA, atpF, atpE, atpI, cemA, ndhA, ndhB, ndhC, ndhf, ndhG, ndhH, ndhJ, ndhE, ndhI, petA, petG, petN, petL, rps8, rps7, rps4, rps14, rps3, rps19, rps2, rps11, rps18, psaB, psaI, psaC, psaJ, psbD, psbI, psbK, psbM, psbA, psbL, psbH,psbC, psbJ, psbT, psbZ, psbK, rpl14, rpl12, rpl20, rpl20, rpl36, rpl33, rpl32* and *rpl23.*

IR regions' expansion and contraction in the chloroplast genome

The study of Cp genomes across different *Salacca* species reveals intriguing patterns of conservation and variation, which are crucial to understanding their evolution and functional dynamics; genes like *rps19, rpl22, ndhF, ycf1, trnH,* and *psbA* are conserved across species, especially at the border regions. As *S. wallichiana is a* partial genome, it shows abnormalities in gene conservation. The *rps3 gene is highly* conserved in all *Salacca* species except *S. wallichiana*. The *rpl22* is conserved completely in LSC regions in *S. zalacca, S. affinis*, and *S. secunda* that are distanced by 31,36,28 bp from the IR_B region, respectively, while in *S. sumatrana* and *S. glabrescens* extend beyond LSC into IR_B by two bp and *S. ramosiana* is inside LSC region. For *rps19*, it is entirely in the IRB region, except for *S. ramosiana*, which is distanced by 18 bp from the LSC region. *trnH* gene in *S. zalacca, S. ramosiana,* and *S. secunda* is present in IR_B region and absent in other species.

The *ycf1* gene in *S. zalacca* and *S. sumatrana* is extended by 4,212 bp and 4,218 bp in the SSC region. The *ycf1* in *S. sumatrana, S. ramosiana,* and *S. glabrescens* near the border of SSC but still inside the IRB region, but for *S. secunda* and *S. wallichiana*, the *ycf1* gene is extended by 9 bp and 102 bp in SSC region respectively. The *ndhF* gene in *S. zalacca* and *S. sumatrana* is extended beyond 55 bp into the IR_A region. The first three species, namely *S. ramosiana, S. affinis*, and *S. galbrescens*, exhibit sequence extensions more significant than 56 bp into the IR_A region from the SSC region, whereas for the fourth one, i.e., *S. secunda*, it is observed to be more than 44 bp and *S. wallichiana* is the distance by nine bp from IR_B region. The *ycf1* gene in *S. ramosiana, S. affinis, S. glabrescens, S. secunda*, and *S. wallichiana* is extended by 1,346 bp for the first three species and 1,367 bp and 1,223 bp beyond the IR_A region from the SSC region. The *rpl2* gene is present in the IR_A region of *S. zalacca, S. ramosiana,* and *S. secunda* only, and the *trnH* gene is present in all other species except *S. wallichiana.* The *rps19* gene for *S. zalacca and S. ramosiana* is distanced by 36 bp and 19 bp from the LSC region, and *S. secunda* is extended 26 bp beyond the LSC region, respectively.

The *psbA* gene for *S. sumatrana*, *S. affinis*, and *S. glabrescens* is 100 bp, 131 bp, and 100 bp beyond the IR_A region, respectively. Other species have the *trnL* gene somewhere in the middle of the LSC region except *S. wallichiana*, which has distanced by 185 bp from the IRA region and has 106,008 bp, which is abnormal from other *Salacca* species (Figure 3). This conservation underscores these gene's critical roles in the Cp genome stability and function. Despite this conservation, Variations in the border regions, including small insertions, deletions, or inversions, highlight potential evolutionary adaptations. These differences may provide information about the evolutionary relevance of these Cp genomes. Typically, the IR region is a primary factor for changes in the size of the Cp genome, leading to expansion, shrinkage, and loss of genetic material (Bock and Knoop 2012).

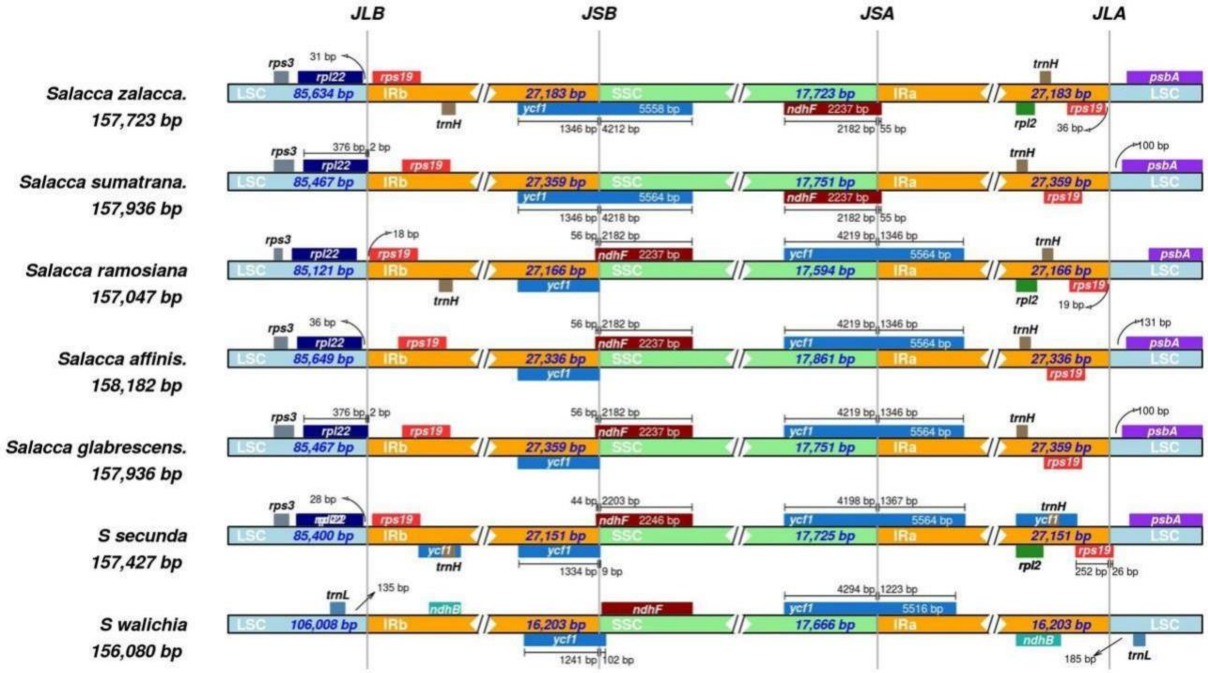

Figure 3. Seven-chloroplast genomes IR/LSC and IR/SSC border locations are compared, including *Salacca zalacca*, *S. sumatrana*, *S. ramosiana*, *S. affinis*, *S. glabrescens*, *S. secunda* and *S. walichiana*

Chloroplast microsatellites: Comparative quantity and distribution among species

There is a remarkable level of conservation in the Cp genomes. Closely related species differ in the number of SNPs, InDels, and chloroplast simple sequence repeats (cpSSRs). SNP is the most prevalent throughout the genome of all the genetic variants. The frequency of InDels is lower than that of SNPs. The presence of repetition and variation sequences in cp genomes supports the hypothesis that the mutation rate in coding regions is less than in the non-coding areas (Niu et al. 2017). In the Cp of angiosperms, LSC and SSC regions have more SSR than IR sections (Gao et al. 2018). For a variety of genetic analyses, such as population genetics, phylogeography, ecological, systematic conservation, and the creation of molecular markers for phylogenetic research, DNA fingerprinting, and plant breeding, genetic variations can be a valuable source of genetic material (Andrade et al. 2018). The *Salaccca* species differ in the total number of cpSSR. Based on the Phobos software's output in Geneious Prime 2019.1.1 version 11 (https://www.geneious.com). *S. zalacca*, *S. affinis*, *S. sumatrana*, *S. ramosiana*, *S. glabrescens*, *S. wallichiana* and *S. secunda* have the total of cpSSRs in their respective cpDNA, i.e. 295, 299, 300, 289, 296, 294 and 301 cpSSRs identified. In the intergenic areas, including cpSSRs, there are 94, 103, 100, 104, 97, 97, and 104 for mono-, 31,34, 34, 33, 33,33, and 29 for di-, 28, 32,31, 35, 30,24, and 29 tri-nucleotide repeats, respectively, while in the genic regions, 76, 70, 71, 66, 72, 72 and 72 for mono-29, 26, 28, 21, 29,26, and 30 for di- and for tri- 37, 34, 34, 30, 35, 42, and 36 cpSSRs are identified, respectively (Table 11). The mononucleotide is the most common SSR motif in the *Salacca* Cp genome, followed by the dinucleotide (Di-). The *psbB, trnK-UUU, rps16, rps19, psbK, trnS-CGA, psbT, atpH, psbC, trnE-UUC rps14, pafI, pafII, rps4, trnE-UUC, cemA, petA, petB* and *psaJ* detected mono-nucleotide motifs in all *Salacca* species. The *rrn23, trnS-GCU, trnS-UGA, trnL-UAA,* and *ndhH* detected mono-nucleotide (Mono-). The trinucleotide (Tri) motif is observed in psbA, rps15, rps18, rpl2, rpl22, rrn16, psaB, ccsA, and *ndhK*. On the other hand, *matK, rpl16*, and *ndhF* contain mono- and tri- motif, while *rpoC2* and *accD* have mono- and di-motifs, and *rpoC1, ndhB, clpP1, ycf1, ycf2*, and *ndhD* from all *Salacca* species that contain all three, Mono-, Di-, Tri-motifs.

Phylogenetic analysis

The phylogenetic tree analysis shows the evolutionary connections among multiple palm species, including *Salacca* varieties, using *Dasypogon bromeliifolious* NC_020367 (Db) as an outgroup. On the tree, it is evident that the *Salacca* species are highly akin to the *Calamus* genus. Together, they form a clade implying more recently shared lineage than any other species. The clade comprising *Salacca* and *Calamus* is a sibling to another containing *Mauritia* and *Eugeissona* genera. The tree of the *Salacca* genus reveals that *Salacca secunda* and *Salacca ramosiana* exhibit the highest level of relatedness among all species. These two closely linked species form a clade adjacent to one another, containing *S. affinis* and *S. sumatrana* as sister groups (Figure 4).

Table 11. The frequency of chloroplast SSR in the coding (Genic CpSSR) and non-coding regions (Intergenic Cp SSR) of seven *Salacca* species

Species	Genic CpSSR			Intergenic CpSSR		
	Mono-	Di-	Tri-	Mono-	Di-	Tri-
S. affinis	70	26	34	103	34	32
S. sumatrana	71	28	34	100	34	31
S. glabrescens	72	29	35	97	33	30
S. zalacca	76	29	37	94	31	28
*S. wallichiana**	72	26	42	97	33	24
S. secunda	73	30	36	104	29	29
S. ramosiana	66	21	30	104	33	35

Note: *Partial Cp genome

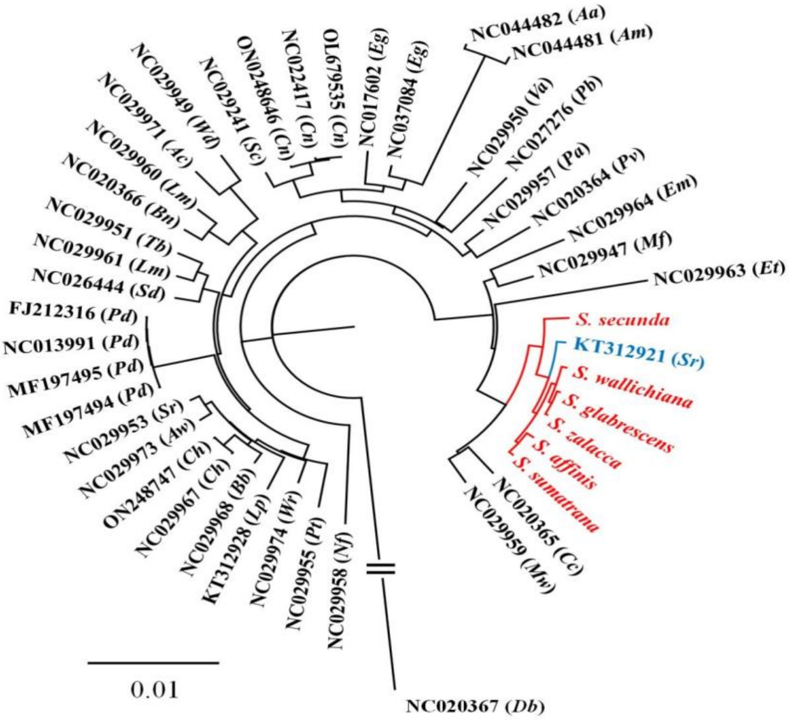

Figure 4. Phylogenetic tree of Cp genome

In comparison, other members of the same family, such as *S. glabrenscens* and *S. wallichiana*, play lesser degrees of association with their counterparts on this phylogenetic map. According to the tree, the *Salacca* genus appears to have a closer genetic connection with the coconut palm (*Cocos nucifera*) than the date palm (*Phoenix dactylifera*). This finding is somewhat unexpected since coconut and date palms belong to the subfamily Arecoideae, while *Salacca* falls under Calamoideae. Revising existing subfamily classifications may be necessary. Valuable insights into the evolutionary relationships of *Salacca* species and other palm trees can be obtained from the phylogenetic tree. Such information can serve as a basis for conservation efforts, breeding programs, and various research endeavors.

In conclusion, this study presents novel and comprehensive Cp genome sequences from six *Salacca* species. Additionally, the Cp genomes were evaluated with other species. The complete Cp genome from six *Salacca* species is successfully assembled. However, only a partial Cp genome of *S. wallichiana* is obtained. The *Salacca* Cp genomes exhibited genome content, structure, and gene order similarities, except for *S. wallichiana*, due to its incomplete Cp genome. An analysis of the seven *Salacca* cp genomes identified 813 SNPs and 369 InDels that can be used to develop molecular markers based on the Cp genome in the future. Such molecular markers could be used to investigate phylogenetics, comparative genomics, and the functional implications of non-synonymous SNPs in genes such as *ycf1*, *ycf2*, and *rpoC2*. Examining the nucleotide content and arrangement of the genome can offer valuable information about the stability of the Cp genome and their gene expression regulation, which fills the gaps in the understanding of plant genetics and facilitates breakthroughs in biotechnology.

ACKNOWLEDGEMENTS

The first author gratefully acknowledges the support of the KNB Scholarship Program and the Ministry of Higher Education, Republic of Indonesia, for the MSc. program. The authors sincerely thank the Plant Molecular Biology Laboratory (PMB Lab), Department of Agronomy and Horticulture, Faculty of Agriculture, Institut Pertanian Bogor, Bogor, West Java, Indonesia, for providing the essential research equipment and the PT. Bumitama Gunajaya Agro (BGA) for partial funding of this work.

REFERENCES

Ahmed I. 2015. Chloroplast genome sequencing: Some reflections. J Next Generat Seq Appl 2: 000119N. DOI: 10.4172/2469-9853.1000119.

Amiryousefi A, Hyvönen J, Poczai P. 2018. IRscope: An online program to visualize the junction sites of chloroplast genomes. Bioinformatics 34 (17): 3030-3031. DOI: 10.1093/bioinformatics/bty220.

Andrade MC, Perek M, Pereira FB, Moro M, Tambarussi EV. 2018. Quantity, organization, and distribution of chloroplast microsatellites in all species of *Eucalyptus* with available plastome sequence. Crop Breed Appl Biotechnol 18: 97-102. DOI: 10.1590/1984-70332018v18n1a13.

Bi Y, Zhang M, Xue J, Dong R, Du Y, Zhang X. 2018. Chloroplast genomic resources for phylogeny and DNA barcoding: A case study on *Fritillaria.* Sci Rep 8: 1184. DOI: 10.1038/s41598-018-19591-9.

Bock R, Knoop V. 2012. Genomics of chloroplasts and mitochondria. Springer, Netherlands. DOI: 10.1007/978-94-007-2920-9.

Chen J, Hao Z, Xu H, Yang L, Liu G, Sheng Y, Zheng C, Zheng W, Cheng T, Shi J. 2015. The complete chloroplast genome sequence of the relict woody plant *Metasequoia glyptostroboides* Hu et Cheng. Front Plant Sci 6: 447. DOI: 10.3389/fpls.2015.00447.

Deng N, Zhou H, Fan H, Yuan Y. 2017. Single nucleotide polymorphisms and cancer susceptibility. Oncotarget 8 (66): 110635-10649. DOI: 10.18632/oncotarget.22372.

Downie SR, Jansen RK. 2015. A comparative analysis of whole plastid genomes from the *Apiales*: Expansion and contraction of the inverted repeat, mitochondrial to plastid transfer of DNA, and identification of highly divergent non-coding regions. Syst Bot 40 (1): 336-351. DOI: 10.1600/036364415X686620.

Gan P, Liu F, Li R, Wang S, Luo J. 2019. Chloroplasts-beyond energy capture and carbon fixation: Tuning of photosynthesis in response to chilling stress. Intl J Mol Sci 20: 5046. DOI: 10.3390/ijms20205046.

Gao X, Zhang X, Meng H, Li J, Zhang D, Liu C. 2018. Comparative chloroplast genomes of Paris Sect. *Marmorata*: Insights into repeat regions and evolutionary implications. BMC Genomics 19 (10): 878. DOI: 10.1186/s12864-018-5281-x.

Green BR. 2011. Chloroplast genomes of photosynthetic eukaryotes. Plant J 66 (1): 34-44. DOI: 10.1111/j.1365-313X.2011.04541.x.

Gun L, Yumiao R, Haixian P, Liang Z. 2018. Comprehensive analysis and comparison on the codon usage pattern of whol*e Mycobacterium tuberculosis* coding genome from different areas. Biomed Res Intl 2018: 3574976. DOI: 10.1155/2018/3574976.

Harnelly E, Thomy Z, Fathiya N. 2018. Phylogenetic analysis of *Dipterocarpaceae* in Ketambe Research Station, Gunung Leuser National Park (Sumatra, Indonesia) based on rbcL and matK genes. Biodiversitas 19 (3): 1074-1080. DOI: 10.13057/biodiv/d190340.

Huang Y, Wang J, Yang Y, Fan C, Chen J. 2017. Phylogenomic analysis and dynamic evolution of chloroplast genomes in Salicaceae. Front Plant Sci 8: 1050. DOI: 10.3389/fpls.2017.01050.

Irwani AN, Oskandar YA, Rahmawati DP, Tara PDR. 2022. Process of making Nata de *Salacca* from honey salak fruit (***Salacca edulis Reinw***) with the application of biotechnology techniques. J Biosci Nat Resour 1 (2): 73-79. DOI: 10.12928/jbns.v1i2.5306.

Ismail NA, Abu Bakar MF. 2018. Salak-*Salacca zalacca*. In: Rodrigues S, de Oliveira Silva E, de Brito ES (eds). Exotic Fruits. Elsevier. DOI: 10.1016/B978-0-12-803138-4.00051-4.

Jansen RK, Ruhlman TA. 2012. Plastid Genomes of Seed Plants. In: Bock R, Knoop V (eds). Genomics of Chloroplasts and Mitochondria. Springer, Netherlands. DOI: 10.1007/978-94-007-2920-9_5.

Jin J-J, Yu W-B, Yang J-B, Song Y, dePamphilis CW, Yi T-S, Li D-Z. 2020. GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol 21 (1): 241. DOI: 10.1186/s13059-020-02154-5.

Kaila T, Chaduvla PK, Rawal HC, Saxena S, Tyagi A, Mithra SVA, Solanke AU, Kalia P, Sharma TR, Singh NK, Gaikwad K. 2017. Chloroplast genome sequence of cluster bean (*Cyamopsis tetragonoloba L.*): Genome structure and comparative analysis. Genes 8 (9): 212. DOI: 10.3390/genes8090212.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol Biol Evol 33 (7): 1870-1874. DOI: 10.1093/molbev/msw054.

Lestari R, Ebert G, Juanda H, Obstbau F. 2002. Salak (*Salacca zalacca* (Gaertner.) Voss.) - The Snakefruit from Indonesia: Preliminary Results of an Ecophysiological Study. https://www.doc-developpement-durable.org/file/Culture/Arbres-Fruitiers/FICHES_ARBRES/salak/Salak%20Salacca%20zalacca%20%E2%80%93%20The%20Snakefruit%20from%20Indonesia.pdf

Li X, Yang Y, Henry RJ, Rossetto M, Wang Y, Chen S. 2015. Plant DNA barcoding: From gene to genome. Biol Rev 90 (1): 157-166. DOI: 10.1111/brv.12104.

Liu H, Lu Y, Lan B, Xu J. 2020. Codon usage bias in chloroplast genes of *Hemiptelea davidii.* J Genet 99 (1): 8. DOI: 10.1007/s12041-019-1167-1.

Lowe TM, Chan PP. 2016. tRNAscan-SE Online: Integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res 44 (W1): W54-W57. DOI: 10.1093/nar/gkw413.

Matra DD, Ritonga AW, Natawijaya A, Poerwanto R, Sobir, Siregar UJ, Widodo WD, Inoue E. 2019. Datasets for genome assembly of six underutilized Indonesian fruits. Data Brief 22: 960-963. DOI: 10.1016/j.dib.2018.12.070.

Menezes APA, Resende-Moreira LC, Buzatti RSO, Nazareno AG, Carlsen M, Lobo FP, Kalapothakis E, Lovato MB. 2018. Chloroplast genomes of *Byrsonima* species (Malpighiaceae): Comparative analysis and screening of high divergence sequences. Sci Rep 8: 2210. DOI: 10.1038/s41598-018-20189-4.

Niu Z, Pan J, Zhu S, Li L, Xue Q, Liu W, Ding X. 2017. Comparative analysis of the complete plastomes of *Apostasia wallichii* and *Neuwiedia singapureana* (*Apostasioideae*) reveals different evolutionary dynamics of ir/ssc boundary among photosynthetic orchids. Front Plant Sci 8: 1713. DOI: 10.3389/fpls.2017.01713.

Rahmawati A, Volkaert HA, Dinarti D, Maskromo I, Hatta ANN L, Sudarsono S. 2021. Complete chloroplast genome sequences of coconut cv. kopyor green dwarf and comparative genome analysis to oil palm, date palm, sago palm, and miniature sugar palm. In: Tombuloglu H, Unver T, Tombuloglu G, Hakeem KR (eds). Oil Crop Genomics. Springer International Publishing. DOI: 10.1007/978-3-030-70420-9_10.

Saleh MSM, Siddiqui M, Mediani A, Ismail N, Ahmed Q, So'ad SM, Saidi-Besbes S. 2018. *Salacca zalacca*: A short review of the palm botany, pharmacological uses and phytochemistry. Asian Pac J Trop Med 11 (12): 645. DOI: 10.4103/1995-7645.248321.

Sehn JK. 2015. Insertions and deletions (Indels). In: Kulkarni S, Pfeifer J. (eds). Clinical Genomics. Elsevier. DOI: 10.1016/B978-0-12-404748-8.00009-5.

Silitonga YW, Lubis RH, Harahap QH. 2019. Utilization of *salak sidimpuan* (*Salacca sumatrana Becc*) as a nata de *Salacca* substrate in Sitaratoit Village South Tapanuli Selatan Sumatera Utara. J Saintech Transfer 1 (2): 175-180. DOI: 10.32734/jst.v1i2.818.

Visendi P, Batley J, Edwards D. 2014. Next-generation sequencing and germplasm resources. In: Tuberosa R, Graner A, Frison E (eds). Genomics of Plant Genetic Resources. Springer Netherlands. DOI: 10.1007/978-94-007-7572-5_15.

Wang Y, Zhan D-F, Jia X, Mei W-L, Dai H-F, Chen X-T, Peng S-Q. 2016. Complete chloroplast genome sequence of *Aquilaria sinensis* (*Lour.*) Gilg and evolution analysis within the Malvales order. Front Plant Sci 7: 280. DOI: 10.3389/fpls.2016.00280.

Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D. 2011. The evolution of the plastid chromosome in land plants: Gene content, gene order, gene function. Plant Mol Biol 76: 273-297. DOI: 10.1007/s11103-011-9762-4.

Wilson CA. 2004. Phylogeny of Iris based on chloroplast matK gene and trnK intron sequence data. Mol Phylogenet Evol 33 (2): 402-412. DOI: 10.1016/j.ympev.2004.06.013.

Yamada KD, Tomii K, Katoh K. 2016. Application of the MAFFT sequence alignment program to large data-reexamination of the usefulness of chained guide trees. Bioinformatics 32 (21): 3246-3251. DOI: 10.1093/bioinformatics/btw412.

Yang M, Zhang X, Liu G, Yin Y, Chen K, Yun Q, Zhao D, Al-Mssallem IS, Yu J. 2010. The complete chloroplast genome sequence of date palm (*Phoenix dactylifera* L.). PLoS One 5 (9): e12762. DOI: 10.1371/journal.pone.0012762.

Yang J, He J, Wang DB, Shi E, Yang W, Geng Q, Wang Z. 2016. Progress in research and application of InDel markers. Biodiver Sci 24 (2): 237-243. DOI: 10.17520/biods.2015205.

Yu X-Q, Drew BT, Yang J-B, Gao L-M, Li D-Z. 2017. Comparative chloroplast genomes of eleven schima (*Theaceae*) species: Insights into DNA barcoding and phylogeny. PLoS One 12 (6): e0178026. DOI: 10.1371/journal.pone.0178026.

Zhang Y, Du L, Liu A, Chen J, Wu L, Hu W, Zhang W, Kim K, Lee S-C, Yang T-J, Wang Y. 2016. The complete chloroplast genome sequences of five *Epimedium* species: Insights into phylogenetic and taxonomic analyses. Front Plant Sci 7: 306. DOI: 10.3389/fpls.2016.00306.

Zumaidar, Miftahuddin. 2018. Species distribution of genus *Salacca*. J Phys Conf Ser 1116: 052083. DOI: 10.1088/1742-6596/1116/5/052083.

Supplementary data:

Genes encoded by the *Salacca sumatrana* Cp genome

Category of genes	Group of gene					
Self-replication	Ribosomal RNA genes	*rrn23s(x2)*	*rrn16s(x2)*	*rrn4.5s(x2)*	*rrn5s(x2)*	
	Transfer RNA genes	*trnV-UAC,*	*trnM-CAU,*	*trnfM-CAU,*	*trnA-UGC(x2)*	
		trnS-GGA,	*trnT-UGU,*	*trnT-GGU,*	*trnV-GAC(x2)*	
		trnG-GCC,	*trnC-GCA,*	*trnS-UGA,*	*trnN-GUU(x2)*	
		trnL-UAA,	*trnI-GAU,*	*trnE-UUC,*	*trnR-ACG(x2)*	
		trnF-GAA,	*trnY-GUA,*	*trnK-UUU*	*trnH-GUG(x2),*	
		trnL-UAG,	*trnD-GUC,*		*trnl-CAU(x2)*	
		trnW-ssCCA,	*trnR-UCU,*			
		trnP-UGG,	*trnG-UCC,*			
		trnL-CAA,	*trnS-GCU,*			
		trnl-GAU	*trnQ-UUG*			
	Small Subunits of ribosomes	*rps2*	*rps3*	*rps4*	*rps7(x2) rps15*	*rps8*
		rps11	*rps12 (x3)*	*rps14*		*rps1*
		rps18	*rps19(x2)*			
	Large Subunit of ribosomes	*rpl 2(x2)*	*rpl14*	*Rpl16*	*rpl20*	*rpl22*
		rpl 23(x2)	*rpl32*	*rpl33*	*rpl36*	
	DNA-dependent RNA polymerase	*rpoA*	*rpoB*	*rpoC1*	*rpoC2*	
	Subunit of NADH-Dehydrogenase	*ndhA*	*ndhB(x2)*	*ndhC, ndhH*	*ndh D,ndhI*	*ndhE,*
		ndhF	*ndhG*			*ndhJ*
		ndhK(x2)				
	Subunit of photosystem I	*psaA*	*psaB*	*psaJ*	*psaI*	*psaC*
	Subunit of photosystem II	*psbA*	*psbB*	*psbC*	*psbD*	*psbE*
		psbF	*psbH*	*psb I*	*psbJ*	*psbK*
		psbL	*psbM*		*psbT*	*psbZ*
	Subunit of Cytochrome b/f complex	*petA*	*petB*	*petD(x2)*	*petL,*	*petN*
						petG
Genes for photosynthesis	Subunit of ATP synthase	*atpA*	*atpB*	*atpE*	*atpF*	*atpH*
		atpI				
	Subunits of rubisco	*Rbcl*				
	Maturase	*matK*				
	Envelope membrane protein	*cemA*				
Others	Subunit of acetyl-CoA Carboxylase	*accD*				
	C-type Cytochrome synthesis gene	*ccsA*				
	Translational initiation factor	*infA*				
Genes of unknown function	Conserved open reading frames	*ycf2(x2)*	*ycf1*	*pafl*	*pbf1*	
				pafll		
	Protease	*clpP1*				