 doi:10.20944/preprints201807.0458.v1

Article

Two-Step Contractions of Inverted Repeat Region and *Psa*i Gene Duplication from the Plastome of *Croton Tiglium* (Euphorbiaceae)

Sangjin Jo and Ki-Joong Kim*

Division of Life Sciences, Korea University, Seoul 02841, Korea; adidevil@korea.ac.kr
* Correspondence: kimkj@korea.ac.kr; Tel.: +82-2-3290-3454

Abstract: *Croton* L. (Euphorbiaceae) is a very specious genus and consists of about 1,250 species, mainly distributed in tropical Asia and China. The first complete plastome sequence from the genus, *Croton tiglium,* is reported in this study (NCBI acc. No. MH394334). The plastome is 150,021 bp in length. The lengths of LSC and SSC are 111,654 bp and 18,167 bp, respectively. However, the length of the IR region is only 10,100 bp and includes only four *rrn* and four *trn* genes, and a small part of the *ycf*1 gene. We propose two-step IR contractions to explain this unique IR region of the *C. tiglium* plastome. First, the IR contracted from rps19-rpl2 to *ycf*2-*trn*L-CAA on the LSC/IRb boundary. Second, the IR contracted from *ycf*2-*trn*L-CAA to *rrn*16-*trn*V-GAC on the LSC/IRa boundary. In addition, duplicated copies of *psa*I genes were discovered in the *C. tiglium* plastome. Both copies were located side by side between *acc*D and *ycf*4 genes, but one copy was pseudogenized because of a five-basepair (TAGCT) insertion in the middle of the gene following frameshift mutation. The plastome contains 112 genes, of which 78 are protein-coding genes, 30 are tRNA genes, and four are rRNA genes. Sixteen genes contain one intron and two genes have two introns. The *inf*A gene is lost. Twelve large repeats were detected in the plastome. All large repeats are located in the LSC region. Also, 272 simple sequence repeats (SSRs) were identified. The penta-SSRs accounted for 45% of total SSRs, followed by mono- (32%), di- (12%), tetra (6%) and tri-SSRs (5%). Most of them were distributed in the large single copy (LSC) region (85%). In addition, 76% of the SSRs were located in the intergenic spacer (IGS). Phylogenetic analysis suggested that *C. tiglium* is a sister group of *Jatropha curcas* with 100% bootstrap support. Seven Euphorbiaceae species formed one clade with 100% bootstrap support.

Keywords: IR contraction; *psa*I gene duplication; Plastome; SSR; *Croton tiglium*; Euphorbiaceae

1. Introduction

The Euphorbiaceae belongs to Malpighiales and is the seventh-largest family in the flowering plant that contains 6,252 species in 209 genera [1-2]. The Euphorbiaceae was classified into five subfamilies at first [3-6], but was later treated to include four subfamilies (Acalyphoideae, Cheilosoideae, Crotonoideae, and Euphorbioideae) and the subfamily Peroideae was treated as Peraceae, which is an independent family, now [7-8].

The genus *Croton* L. is the second-largest genus after *Euphorbia* in the Euphorbiaceae, comprising more than 1,250 species [9-10]. This genus belongs to the subfamily Crotonoideae [3-6]. Webster (1993) classified *Croton* into 40 sections [11]. Van Ee et al. (2011) divided the genus *Croton* into four subgenera, *Adenophylli*, *Croton*, *Geiseleria,* and *Quadrilobi* [12]. *Croton* is distributed all over the world, ranging from America to Africa, Asia, and Oceania, but particularly many species of it are intensively distributed in southern China and Southeast Asia [13]. Various species of *Croton* are used for wound healing, rheumatism treatment, cancer treatment, or are widely cultivated as ornamental plants [14-15].

Croton tiglium L. is known as Purging Croton, and originated in tropical Asia and China [15]. *C. tiglium* belongs to the subgenus *Croton* and the section *Tiglium* [11-12]. As with other *Croton* species,

C. tiglium is used to treat diseases such as rheumatism and cancer, and is also used as a laxative in China [14-15].

With the development of genome sequencing technology, studies on the evolution of plants using complete plastomes have been actively in progress. Currently, over 2,200 complete plastome sequences can be downloaded from the NCBI database. Among them, in the case of Malpighiales to which Euphorbiaceae belongs, the plastomes of a total of 105 species in nine families have been reported (retrieved July 10, 2018). However, Chrysobalanaceae with 50 species and Salicaceae with 41 species account for most of the foregoing species [16-18]. Also, the plastomes of 14 species of Passifloraceae, three species of Malpighiaceae, one species of Clusiaceae, one species of Erythroxylaceae, one species of Linaceae, and one species of Violaceae have been reported [19-24]. In the case of Euphorbiaceae, to which *Croton* belongs, the plastomes of six species have been reported. These species are classified by subfamily as one species of Acalyphoideae (*Ricinus communis)*, four species of Crotonoideae (*Hevea brasiliensis, Jatropha curcas* and *Manihot esculenta*) and one species of Euphorbioideae (*Euphorbia esula*) [25-30]. There is no study finding yet for the plastomes of *Croton*, which is a large genus consisting of 1,250 species. The plastome structure of Euphorbiaceae, published to date, is almost identical to that of the general angiosperm plastomes. However, a 30 kb inversion has been reported to be present in *Hevea brasiliensis* [28]. Given that the complete plastomes of only six species, which are around 0.1% of the 6,252 species of Euphorbiaceae, have been reported, if more species are studied hereafter, many variations can be discovered.

Therefore, this study reports complete plastomes of the first *Croton* species, which has many species in the genus. Through this study, it was proved that the IR region in the plastome of C. *tiglium* was shortened by two-step contractions. In addition, the amounts and distributions of large repeats were compared in seven Euporbiaceae plastomes. We found *psa*I gene duplication and pseudogenization in the *C. tiglium* plastome and report large amounts of simple sequence repeats (SSRs). These results are expected to be widely used for the development of molecular identification markers for *Croton* species and for studies of Euphorbiace plastome evolution.

2. Results and Discussion

2.1. Structure and Gene Contents of Croton tiglium Complete Plastome

The plastomes of *C. tiglium* were sequenced by Illumina HiSeq 2000 platform. The number of total reads is 26,602,848 (Table 1). Among them, the number of plastome reads is 1,396,325 (5.25%).

Table 1. The statistical data of *Croton tiglium* plastome sequenced in this study.

Species	***Croton tiglium***
No. of total reads	26,602,848
No. of plastome reads (Plastome %)	1,396,325 (5.25%)
Average read length per reads	90 bp
Average coverage	837.7x
Plastome length (bp)	150,021
LSC (bp)	111,654
SSC (bp)	18,167
IR (bp)	10,100
A-T content (%)	63.8

The average of coverage is 837.7x. Therefore, the complete plastome of *C. tiglium* is highly accurate. The plastome of *C. tiglium* show a quadripartite structure similar to the plastome structure

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 July 2018 doi:10.20944/preprints201807.0458.v1

of common flowering plants (Figure 1). The plastome is 150,021 bp in length. The lengths of the LSC, SSC and IR regions are 111,654 bp, 18,167 bp and 10,100 bp, respectively (Table 1). In previous studies, the lengths of Euphorbiaceae plastome ranged from 160,512 bp (*Euphorbia esula*) to 163,856 bp (*Jatropha curcas*) (Figure 2). The length of *C. tigilum* plastome was shorter by around 10 ~ 13kb compared to them. This is because the length of the IR region was shortened and the length of the LSC region was increased (Figures 1 and 2). Despite the length of the *C. tiglium* plastome being shorter than that of other species, the length of SSC is around 1 kb longer compared to *E. esula* and *Jatropha curcas*. The average A-T content of the plastome is 63.8%, whereas that in the LSC, SSC and IR regions is 66.1%, 70.5% and 50.9%, respectively.

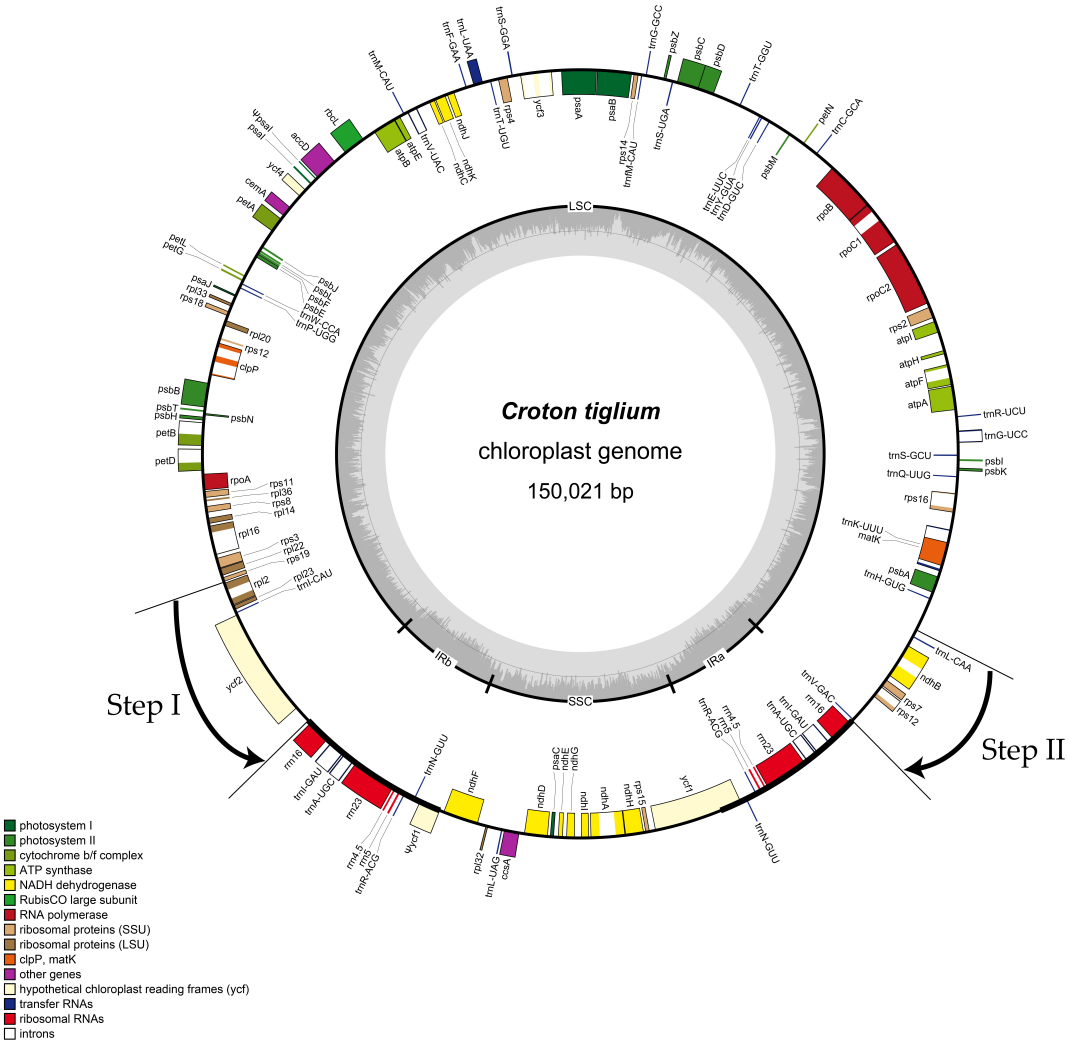

Figure 1. A circular map of *Croton tiglium* (Genbank No. MH394334) plastome. The map was created using OrganellarGenomeDRAW. The pair of thick lines on the outside circle indicate inverted repeats (IRa and IRb). Genes drawn inside the circle are transcribed clockwise, while those drawn outside the circle are transcribed counterclockwise. Step I IR contraction occurred between *rps*19-*rpl*2 and *ycf*2-*trn*L-CAA. Step II IR contraction is located between *ycf*2-*trn*L-CAA and *trn*V-GAC-*rrn*16. Ψ indicates pseudogene.

The plastome comprises 112 unique genes (78 protein-coding genes, 30 tRNA genes and four rRNA genes). Four tRNA and four rRNA genes are duplicated in the IR regions (Figure 1, Table 2).

 doi:10.20944/preprints201807.0458.v1

Protein-coding genes only appeared in the LSC and SSC regions. The 16 genes have one or two introns (Table 2). Two genes (*clp*P and *ycf*3 genes) have two introns. The others have one intron. Both the *psa*I and the Ψ*psa*I are present between *acc*D and *ycf*4. Ψ*psa*I was pseudogenized as TAGCT was inserted (Figure 3). In the six previously reported Euphorbiaceae plastomes, only *psa*I exists between *acc*D and *ycf*4 [25-30]. *Jatropha curcas* is 9 bp longer than other species. The reason for this is that whereas the poly A of other species is 7 bp, that of *Jatropha curcas* became 6 bp, leading to an increase in the length of *psa*I (Figure 2). The *inf*A gene was lost in *C. tiglium*, which is identical to the results of previous Euphorbiaceae plastome studies.

Table 2. Gene contents in *Croton tiglium* plastome sequenced in this study.

Category for genes	Group of genes	Name of genes
Self-replication	rRNA genes	*rrn*16 (x2), *rrn*23 (x2), *rrn*4.5 (x2), *rrn*5 (x2)
	tRNA genes	30 *trn* genes (4 genes are in IR regions)
	Small subunit of ribosome	*rps*2, *rps*3, *rps*4, *rps*7, *rps*8, *rps*11, *rps*12*, *rps*14, *rps*15, *rps*16*, *rps*18, *rps*19
	Large subunit of ribosome	*rpl*2*, *rpl*14, *rpl*16*, *rpl*20, *rpl*22, *rpl*23, *rpl*32, *rpl*33, *rpl*36
	DNA dependent RNA polymerase	*rpo*A, *rpo*B, *rpo*C1*, *rpo*C2
Genes for photosynthesis	Subunits of NADH-dehydrogenase	*ndh*A*, *ndh*B*, *ndh*C, *ndh*D, *ndh*E, *ndh*F, *ndh*G, *ndh*H, *ndh*I, *ndh*J, *ndh*K
	Subunits of photosystem 1	*psa*A, *psa*B, *psa*C, *psa*I, Ψ*psa*J
	Subunits of photosystem 2	*psb*A, *psb*B, *psb*C, *psb*D, *psb*E, *psb*F, *psb*H, *psb*I, *psb*J, *psb*K, *psb*L, *psb*M, *psb*N, *psb*T, *psb*Z
	Subunits of cytochrome b/f complex	*pet*A, *pet*B*, *pet*D*, *pet*G, *pet*L, *pet*N
	Subunits of ATP synthase	*atp*A, *atp*B, *atp*E, *atp*F*, *atp*H, *atp*I
	Large subunit of rubisco	*rbc*L
Other genes	Translational initiation factor	
	Maturase	*mat*K
	Protease	*clp*P**
	Envelope membrane protein	*cem*A
	Subunit of Acetyl-CoA-carboxylase	*acc*D
	c-type cytochrome synthesis gene	*ccs*A
Genes of unknown functions	Open Reading Frames (ORF, ycf)	*ycf*1, *ycf*2, *ycf*3**, *ycf*4

One and two asterisks indicate one- and two-intron containing genes, respectively. Genes located on the IR region indicate by the (x2) symbol after gene name.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 July 2018 doi:10.20944/preprints201807.0458.v1

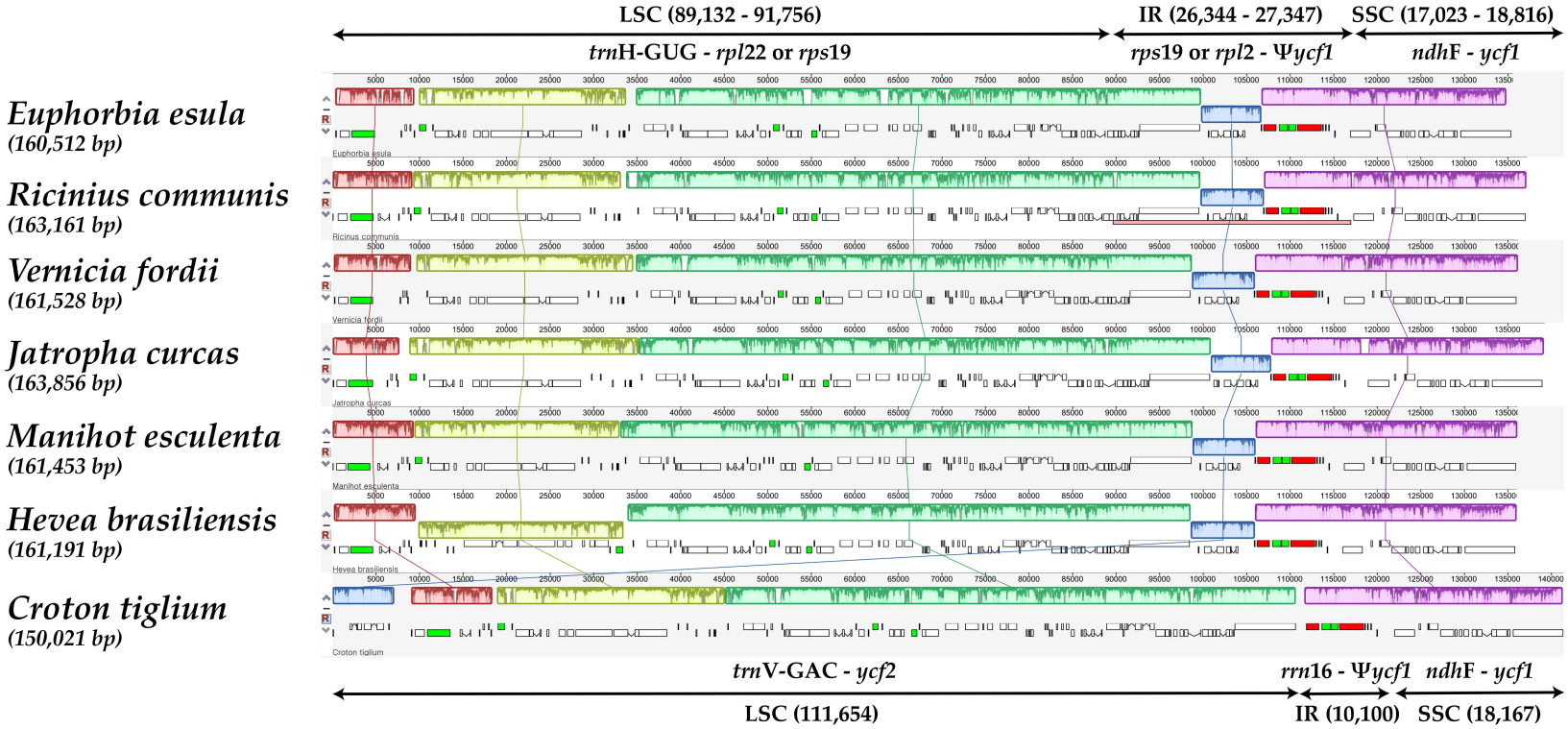

Figure 2. Comparison of seven plastomes aligned using Mauve software. The blue squares are located between *rrn*16 and *trn*V-GAC at one end and between *ycf*2 and the *trn*L-CAA region at the other end. The IR region of *C. tiglium* plastome is substantially shorter than that of the others.

 doi:10.20944/preprints201807.0458.v1

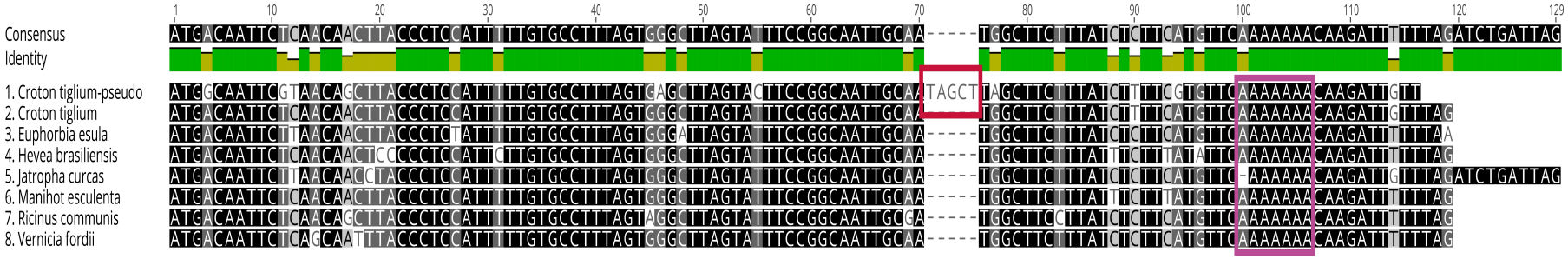

Figure 3. Comparative nucleotide sequences of *psa*I gene among seven Euphorbiaceae plastomes. Two copies of the *psa*I gene occur side by side. The red box indicates the difference between *psa*I and pseudogenized *psa*I. Pseudogenized *psa*I has five-basepair insertion in the middle of the gene and induced the frameshift mutation. The pink box indicates the difference between *Jatropha curcas* and other species.

2.2. Comparison of Plastome Structures in Euphorbiaceae

We compared the plastome structures among seven plastomes in Euphorbiaceae. Unlike six other complete plastomes, the length of the IR region of *C. tiglium* plastome was only 10,100 bp and it included only four *rrn* and four *trn* genes, and a small part of the *ycf*1 gene. The IR region in other Euphorbiaceae are usually more than 26,000 bp in length (Figure 2). In order to explain the current short IR in the *C. tiglium* plastome, we propose a two-step contraction mechanism based on comparative analyses with other plastomes (Figure 1). The *rps*19-*rrn*16 region is generally located at IR in angiosperms [31-33], and in other Euphorbiaceae plastomes too (Table 3). However, it is located in the LSC region in *C. tiglium*. This is because the first IR contraction occurred in the region from *rps*19-*rpl*2 to *ycf*2-*trn*L-CAA and then the second IR contraction occurred in the region from *ycf*2-*trn*L-CAA to *rrn*16-*trn*V-GAC (Figure 1). As a result, the *rps*19-*trn*L-CAA region is located at the 3'-LSC and the *ycf*2-*rrn*16 region is located at the 5'-LSC (Figures 1 and 2). These forms of IR contractions are reported for the first time. More studies are necessary on whether these IR contractions are a phenomenon occurring in certain taxa in the genus *Croton* or are observed in other genera as well.

The large inversion of 30 kb previously reported in the LSC region (*trn*S-GCU-*trn*T-GGU) of *Hevea brasiliensis* [28] was identified as not being present in *Croton*.

Table 3. *Comparative Characteristics of IR Region in Euphorbiaceae.*

	Croton tiglium	*Euphorbia esula*	*Hevea brasiliensis*	*Vernicia fordii*	*Manihot esculenta*	*Jatropha curcas*	*Ricinus communis*
IR length	10,100	26,344	26,810	26,819	26,954	27,124	27,347
***rps*19**	-	-	Δ	O	Δ	-	O
***rpl*2**	-	O	O	O	O	O	O
***rpl*23**	-	O	O	O	O	O	O
***trn*I-CAU**	-	O	O	O	O	O	O
***ycf*2**	-	O	O	O	O	O	O
***trn*L-CAA**	-	O	O	O	O	O	O
***ndh*B**	-	O	O	O	O	O	O
***rps*7**	-	O	O	O	O	O	O
***rps*12**	-	O	O	O	O	O	O
***trn*V-GAC**	-	O	O	O	O	O	O
***rrn*16**	-	O	O	O	O	O	O
***trn*I-GAU**	O	O	O	O	O	O	O
***trn*A-UGC**	O	O	O	O	O	O	O
***rrn*23**	O	O	O	O	O	O	O
***rrn*4.5**	O	O	O	O	O	O	O
***rrn*5**	O	O	O	O	O	O	O
***trn*R-ACG**	O	O	O	O	O	O	O
***trn*N-GUU**	O	O	O	O	O	O	O

Δ: LSC-IR boundary was located in the *rps*19 gene.

 doi:10.20944/preprints201807.0458.v1

2.3. Distribution of Large and Simple Sequence Repeats

The presence of dispersed repeat sequences in plant DNA has been widely used in comparative analysis between inter- or infra-specific species [34-41]. To identify dispersed large repeats, which are at least 26 bp long and with similarity of 100%, the REPuter program [42] and the find repeats of Geneious were used. As a result, a total of 12 large repeats were found (Table 4). Among them, forward repeats occupied the majority with nine, and the numbers of reverse and palindromic repeats were one and two, respectively. The large repeats of seven Euphorbiaceae complete plastomes, including *C. tiglium*, were compared (Figure 4).

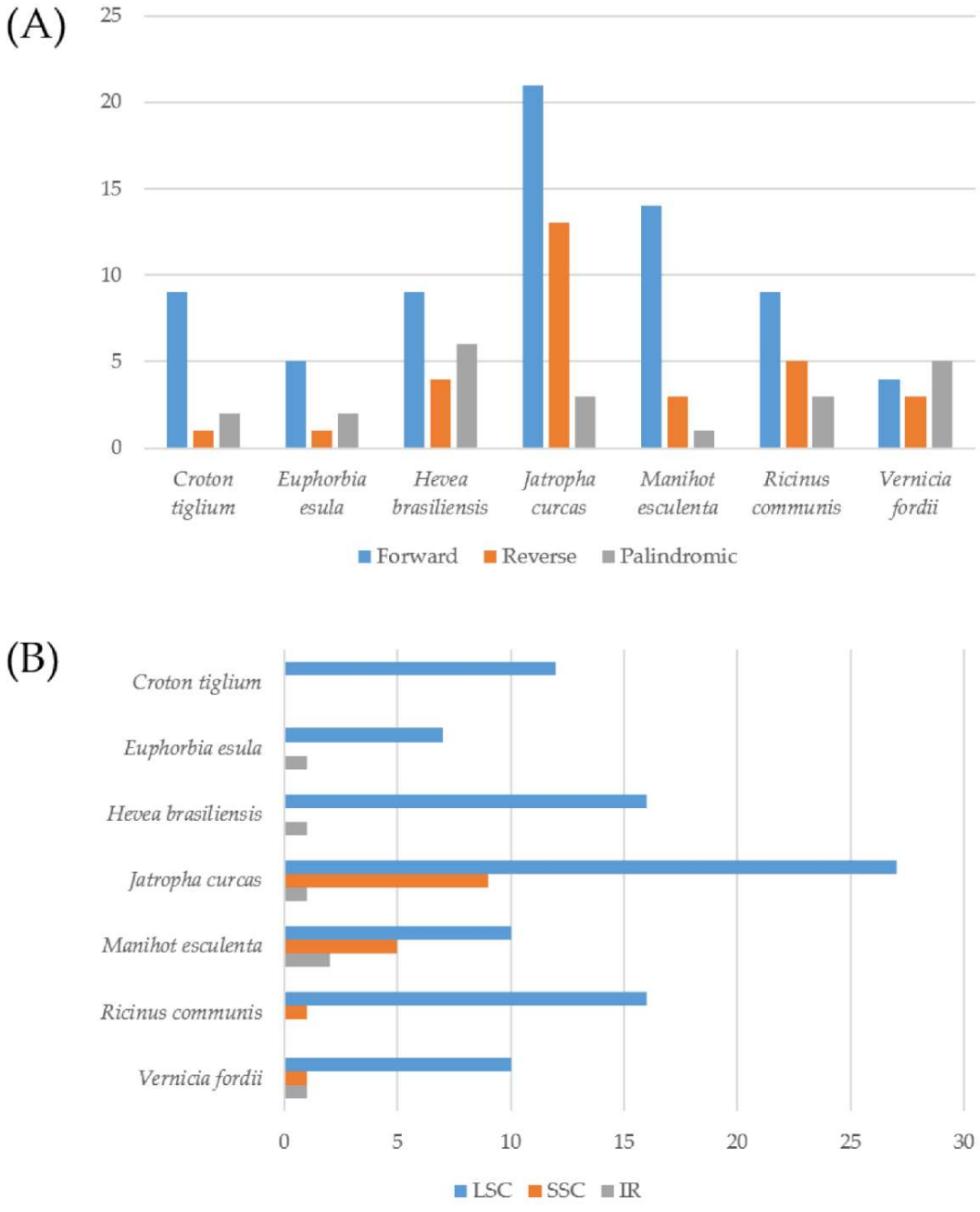

Figure 4. Comparison of large repeats in seven complete plastomes. (A) The number of large repeats (forward, reverse and palindromic repeats). (B) Distribution of large repeats (LSC, SSC and IR regions).

 doi:10.20944/preprints201807.0458.v1

Table 4. Distribution of large repeat loci in *Croton tiglium* plastome.

No.	Size (bp)	Repeat	Location
1	45	Forward	*trn*S-GCU-*trn*G-UCC
2	38	Palindromic	*pet*N-*psb*M
3	36	Forward	*psb*M-*trn*D-GUC / *trn*D-GUC-*trn*Y-GUA
4	35	Forward	*trn*R-UCU-*atp*A
5	33	Forward	*rpo*B-*trn*C-GCA
6	32	Forward	*psb*Z-*trn*G-GCC
7	29	Forward	*psb*E-*pet*L
8	29	Palindromic	*trn*S-GCU / *trn*S-GGA
9	28	Forward	*ycf*3 intron 2
10	27	Reverse	*rps*16 intron
11	26	Forward	*psb*Z-*trn*G-GCC
12	26	Forward	*trn*S-GCU-*trn*G-UCC

In the results, the number of large repeats of *Jatropha curcas* was the largest with 37 and the number of large repeats of *Euphorbia esula* was the smallest with eight (Figure 4A). In all seven species, most of the large repeats were located in the LSC, and in particular, unlike other species, in the case of *C. tiglium*, all large repeats were present in the LSC (Figure 4B). The large repeats of *Euphorbia esula* and *Hevea brasiliensis* were not present in the SSC. In the case of *Ricinus communis* no large repeat was present in the IR.

With regard to simple sequence repeats (SSRs), mono-, di-, tri-, tetra-, and penta-SSRs longer than 10 bp were found using Geneious's Phobos program (Figure 5 and Table S1). A total of 272 SSRs of C. *tiglium* were found, and this number is two to three times larger than the SSRs found in other plant species. Among them, the numbers of penta-SSRs and mono-SSRs were 123 and 87, respectively, and accounted for 77% of the entire SSRs (Figure 5A). These results are in sharp contrast to previous results regarding other taxa, in which mono-SSRs were the most abundant and penta-SSRs were rarely present [19, 41, 43-47]. The largest number of SSRs was located in the LSC (232), followed by the SSC (37) and the IR (3) (Figure 4B). This is because the length of the LSC is as long as 111,654 bp and the rate of variations is high. Unlike the length of a general angiosperm LSC, which is around 85 kb, *C. tiglium* has a larger LSC due to the two-step IR contractions, so that the distribution ratios of SSRs increased in the LSC and decreased in the IR. Most SSRs were distributed in the IGS and 34 and 32 SSRs were distributed in the intron and the CDS, respectively (Figure 5C). This is one of the reasons why the IGS has higher sequence divergences than the CDS. These SSR loci may be usefully used in interspecific or intergroup comparative studies.

2.4. Phylogenetic Analysis

The phylogenetic position of *C. tiglium* was identified using 28 complete plastome sequences belonging to the Fabids (Figure 6). *C. tiglium* formed a clade with *Jatropha curcas* (BS = 100%) and *Vernicia fordii* were located as a sister group of them (BS = 100%). They share inaperturated seed trait and belong to the subfamily Crotonoideae [5-6]. *Manihot esculenta* and *Hevea brasiliensis* form a clade (BS = 100%) in our tree. They formed a sister clade of the *Croton-Jatropha-Vernicia* clade. They also belong to the subfamily Crotonoideae but they share articulated seed traits [5-6]. *Euphorbia esula*, which belongs to the subfamily Euphorbioideae, was located as a sister group of the *Manihot-Hevea* clade (BS = 85%).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 July 2018 doi:10.20944/preprints201807.0458.v1

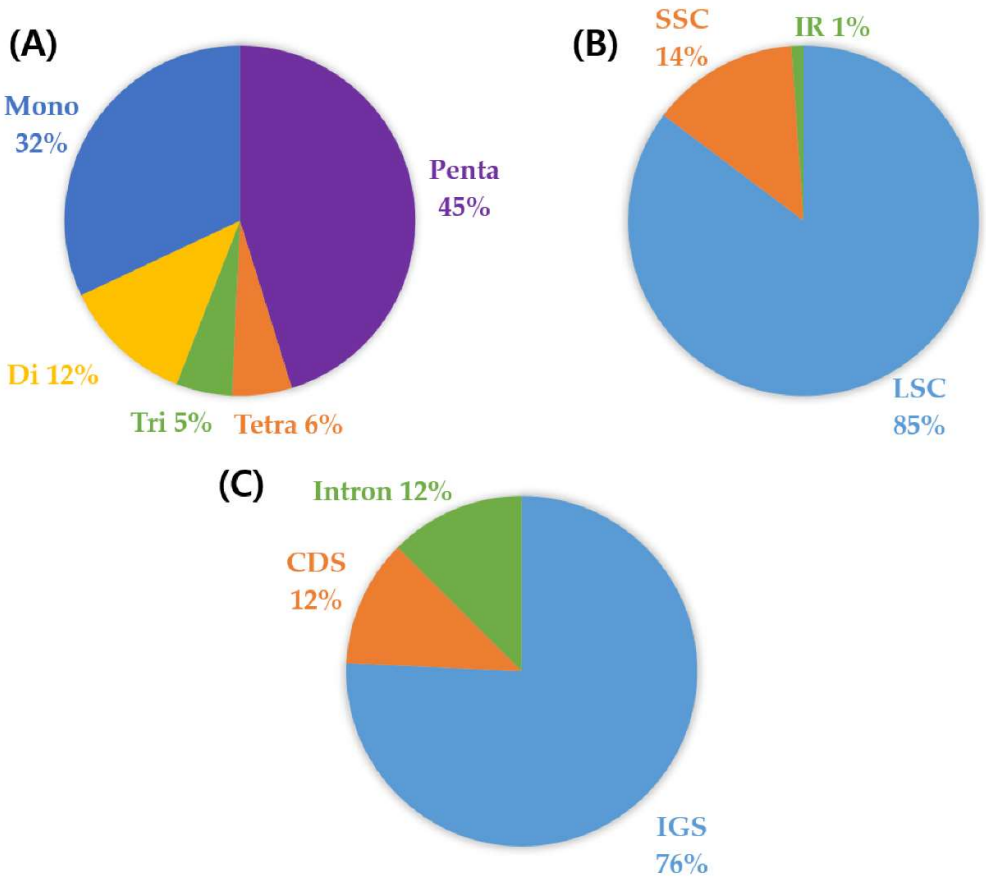

Figure 5. Distribution of simple sequence repeats (SSRs) in *C. tiglium* plastome. (A) Proportion of different SSR types. (B) Proportion of SSRs in LSC, SSC and IR regions. (C) Proportion of SSRs in CDS, IGS and introns.

Our result indicate the subfamily Crotonoideae is paraphyletic group because it included the subfamily Euphorbioideae. *Ricinus communis*, which belongs to Acalyphoideae was located as a basal clade in Euphorbiaceae (BS = 84%). This result is different from the findings in previous studies indicating that Euphorbioideae and Acalyphoideae form sister groups with each other, and Crotonoideae is a sister group of this clade [5-6]. If more whole plastome sequences of Euphorbioideae and Acalyphoideae are added hereafter, accurate results for this difference will be derived.

The results of this study indicate that Euphorbiaceae is a sister group of the (((Salicaceae, Passifloraceae), Violaceae), Linaceae) clade. However, their relationships are uncertain because there is no information on the plastomes of Rafflesiaceae and Peraceae [7-8, 48], which are thought to be closely related to Euphorbiaceae. Hereafter, clear relationships will be established if the whole plastome sequences of Rafflesiaceae and Peraceae are added.

Phylogenetic studies of *Croton* have been conducted several times using partial sequences such as *mat*K, *rbc*L, *trn*L-F, and ITS [5, 49-52]. Phylogenetic trees made using *rbc*L had poor resolutions of *Croton*. However, monophyly was formed in the phylogenetic tree made using *trn*L-F (BS = 100%). Monophyly was also formed along with 100% support in the phylogenetic tree made using combined *rbc*L and *trn*L-F data. In *Croton*'s biogeography studies, C. *tiglium* formed a clade with *C. acutifolius*, *C. argyratus*, *C. cascarilloides*, and *C. megalobotrys* [52].

The first results on complete plastomes of *Croton* obtained in this study will be used as basic data that can be widely applied to *Croton*'s phylogenetic and biogeographic studies hereafter.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 July 2018 doi:10.20944/preprints201807.0458.v1

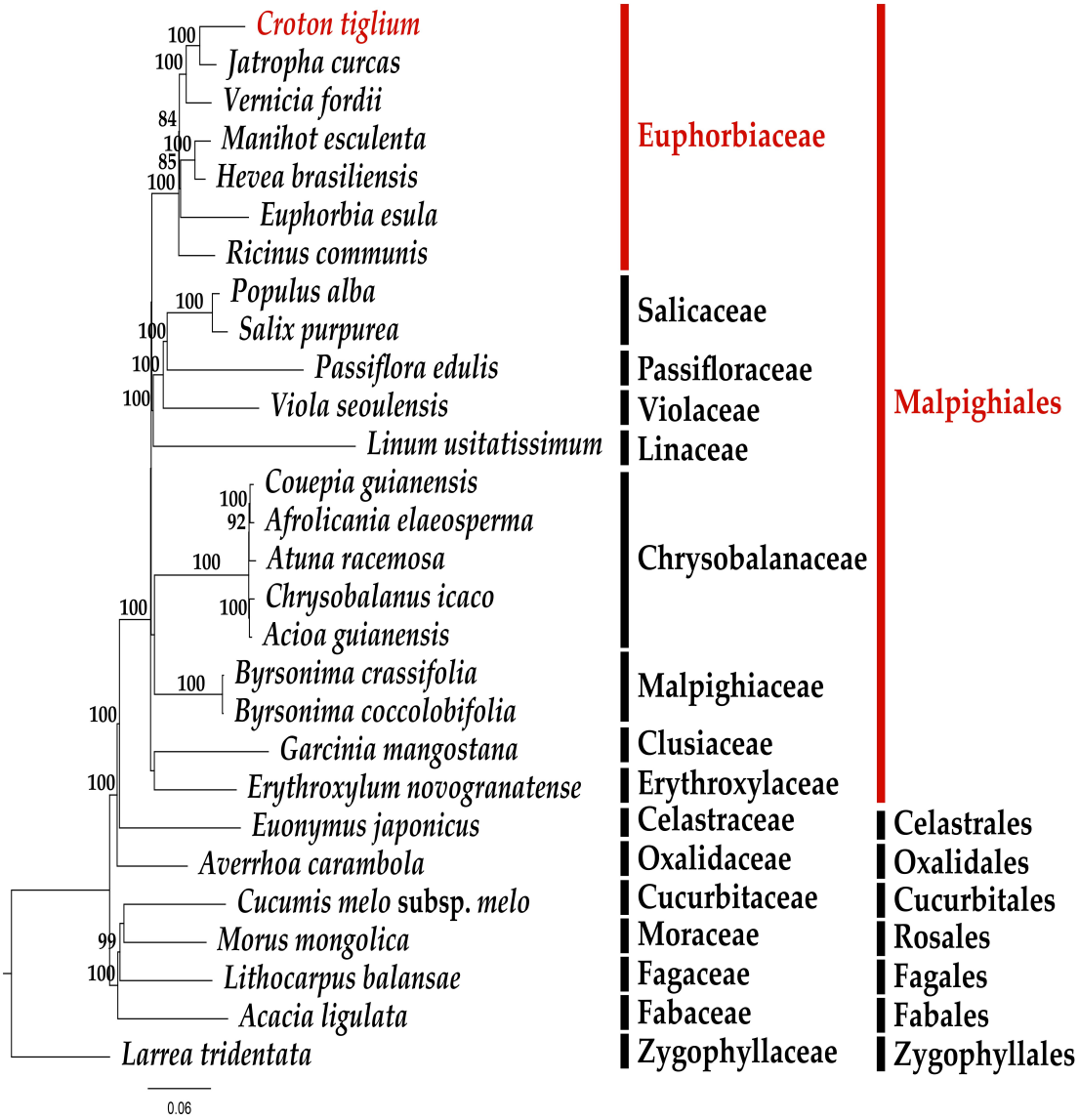

Figure 6. Distribution of simple sequence repeats (SSRs) in *C. tiglium* plastome. (A) Proportion of different SSR types. (B) Proportion of SSRs in LSC, SSC and IR regions. (C) Proportion of SSRs in CDS, IGS and introns.

3. Materials and Methods

3.1. DNA extraction, Sequencing and Annotation

The leaves of *C. tiglium* used in this study were collected from the Korea University greenhouse, where we grew the plants from seeds that were originally collected in Indonesia. The plants flowered and fruited in the greenhouse. A voucher specimen was deposited in the Korea University Herbarium (KUS acc. no. 2014-0242). Fresh leaves were ground into powder in liquid nitrogen and total DNAs were extracted using the CTAB method [53]. The DNAs were further purified by ultracentrifugation and dialysis [54].

The genomic DNAs were deposited in the Plant DNA Bank in Korea (PDBK) under the accession number PDBK2014-0242). The complete plastome sequence was generated using an Illumina HiSeq 2000 system (Illumina, Inc., San Diego, CA). Annotations were performed using the National Center for Biotechnology Information (NCBI) BLAST and tRNAscan-SE programs [55]. The circular map was drawn using the OGDRAW program [56].

3.2. Comparison of structural differences in the Euphorbiaceae plastomes

The Mauve alignment of Geneious was used to compare *C. tiglium* with six Euphorbiaceae plastomes [57]. In addition, MUSCLE alignments with the *psa*I genes of six Euphorbiaceae were carried out to identify the *psa*I and Ψ*psa*I genes existing between *acc*D-*ycf*4 in C. *tiglium* [58].

3.3. Repeat Analysis in the C. tiglium Plastomes

Large repeats in *C. tiglium* were analyzed using the REPuter [42] and find repeats by Geneious. Forward, reverse and palindromic repeats were identified. In addition, we confirmed the large repeats in the six Euphorbiaceae plastomes. For all repeat types, the minimal size and similarity were 26 bp and 100%, respectively.

Simple sequence repeats (SSRs) were identified using Phobos v.3.3.12 from Geneious [60]. The detected SSRs were five types (mono-, di-, tri-, tetra- and penta-SSRs) more than 10 bp.

3.4. Phylogenetic Analysis

A phylogenetic tree was fabricated using whole plastome sequences. To this end, the whole-plastome sequences of 27 species belonging to fabids were downloaded from NCBI (Table S2). Seventy-six protein-coding and four rRNA genes were used in 28 species including C. *tiglium*. *inf*A, *rpl*32, and *rps*16 genes were excluded from the analysis. Each gene was aligned using the MUSCLE program [58] and all genes were concatenated. Maximum likelihood (ML) analysis of the concatenated sequences (79,798 bp) was conducted using RAxML v 7.7.1 [60].

Author Contributions: Ki-Joong Kim conceived and designed the experiments, and collected the materials; Sangjin Jo performed the experiments and wrote the draft of the manuscript; Ki-Joong Kim revised the manuscript. All authors reviewed and approved the final manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) under grant nos. NRF-2015M3A9B8030588 and NRF-2015M3A9B8047398 to KJK.

Acknowledgments: We thank the curator of the Korea University Herbarium (KUS) for voucher specimen preparations. The genomic and chloroplast DNAs are deposited in the Plant DNA Bank in Korea (PDBK).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

LSC	Large Single Copy
SSC	Small Single Copy
IR	Inverted Repeat
IGS	Intergenic Spacer

References

1. APG IV. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot J Linn Soc.* 2016, *181*, 1-20.
2. Christenhusz, M.J.; Byng, J.W. The number of known plants species in the world and its annual increase. *Phytotaxa* **2016**, *261*, 201-217.
3. Webster, G.L. Classification of the Euphorbiaceae. *Ann Mo Bot Gard.* **1994a**, 3-32.
4. Webster, G.L. Synopsis of the genera and suprageneric taxa of Euphorbiaceae. *Ann Mo Bot Gard.* **1994b**, 33-144.
5. Wurdack, K.J.; Hoffmann, P.; Chase, M.W. Molecular phylogenetic analysis of uniovulate Euphorbiaceae (Euphorbiaceae sensu stricto) using plastid *rbc*L and *trn*L-F DNA sequences. *Am J Bot.* **2005**, *92*, 1397-1420.
6. Tokuoka, T. Molecular phylogenetic analysis of Euphorbiaceae sensu stricto based on plastid and nuclear DNA sequences and ovule and seed character evolution. *J Plant Res.* **2007**, *120*, 511-522.

7. Wurdack, K.J.; Davis, C.C. Malpighiales phylogenetics: Gaining ground on one of the most recalcitrant clades in the angiosperm tree of life. *Am J Bot.* **2009**, *96*, 1551-1570.
8. Xi, Z.; Ruhfel, B.R.; Schaefer, H.; Amorim, A.M.; Sugumaran, M.; Wurdack, K.J.; Endress, P.K.; Matthews, M.L.; Stevens, P.F.; Mathews, S. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proc Natl Acad Sci U.S.A.* **2012**, *109*, 17519-17524.
9. Govaerts, R.; Frodin, D.G.; Radcliffe-Smith, A. *World checklist and bibliography of Euphorbiaceae (with Pandaceae)*; Royal Botanic Gardens, Kew, UK, 2000; Volume 4.
10. Frodin, D.G. History and concepts of big plant genera. *Taxon* **2004**, *53*, 753-776.
11. Webster, G.L. A provisional synopsis of the sections of the genus *Croton* (Euphorbiaceae). *Taxon* **1993**, 793-823.
12. van Ee, B.W.; Riina, R.; Berry, P.E. A revised infrageneric classification and molecular phylogeny of new world *Croton* (Euphorbiaceae). *Taxon* **2011**, *60*, 791-823.
13. van Ee, B.W.; Berry, P.E. *Croton* Linnaeus. In *Flora of North America;* Flora of North America Editorial Committee; Oxford University Press: New York, NY, USA, 2016; Volume 12.
14. Salatino, A.; Salatino, M.L.F.; Negri, G. Traditional uses, chemistry and pharmacology of *Croton* species (Euphorbiaceae). *J Braz Chem Soc.* **2007**, *18*, 11-33.
15. Loureiro T. *Croton* Linnaeus. In *Flora of China;* Zhengyi, W., Raven, P.H., Deyuan, H., Eds.; Science Press: Beijing, China; Missouri Botanical Garden Press: St. Louis, MO, USA, 2008; Volume 11, pp. 258-264.
16. Huang, D.I.; Hefer, C.A.; Kolosova, N.; Douglas, C.J.; Cronk, Q.C. Whole plastome sequencing reveals deep plastid divergence and cytonuclear discordance between closely related balsam poplars, *Populus balsamifera* and *P. trichocarpa* (Salicaceae). *New Phytol.* **2014**, *204*, 693-703.
17. Bardon, L.; Sothers, C.; Prance, G.T.; Malé, P.-J.G.; Xi, Z.; Davis, C.C.; Murienne, J.; García-Villacorta, R.; Coissac, E.; Lavergne, S. Unraveling the biogeographical history of Chrysobalanaceae from plastid genomes. *Am J Bot.* **2016**, *103*, 1089-1102.
18. Huang, Y.; Wang, J.; Yang, Y.; Fan, C.; Chen, J. Phylogenomic analysis and dynamic evolution of chloroplast genomes in Salicaceae. *Front Plant Sci.* **2017**, *8*, 1050.
19. Cauz-Santos, L.A.; Munhoz, C.F.; Rodde, N.; Cauet, S.; Santos, A.A.; Penha, H.A.; Dornelas, M.C.; Varani, A.M.; Oliveira, G.C.; Bergès, H. The chloroplast genome of *Passiflora edulis* (Passifloraceae) assembled from long sequence reads: Structural organization and phylogenomic studies in Malpighiales. *Front Plant Sci.* **2017**, *8*, 334.
20. Cheon, K.-S.; Yang, J.-C.; Kim, K.-A.; Jang, S.-K.; Yoo, K.-O. The first complete chloroplast genome sequence from Violaceae (*Viola seoulensis*). *Mitochondrial DNA Part A* **2017**, *28*, 67-68.
21. Jo, S.; Kim, H.-W.; Kim, Y.-K.; Sohn, J.-Y.; Cheon, S.-H.; Kim, K.-J. The complete plastome of tropical fruit *Garcinia mangostana* (Clusiaceae). *Mitochondrial DNA Part B* **2017**, *2*, 722-724.
22. de Santana Lopes, A.; Pacheco, T.G.; dos Santos, K.G.; do Nascimento Vieira, L.; Guerra, M.P.; Nodari, R.O.; de Souza, E.M.; de Oliveira Pedrosa, F.; Rogalski, M. The *Linum usitatissimum* L. plastome reveals atypical structural evolution, new editing sites, and the phylogenetic position of Linaceae within Malpighiales. *Plant Cell Reports* **2018**, *37*, 307-328.
23. Menezes, A.P.; Resende-Moreira, L.C.; Buzatti, R.S.; Nazareno, A.G.; Carlsen, M.; Lobo, F.P.; Kalapothakis, E.; Lovato, M.B. Chloroplast genomes of *Byrsonima* species (Malpighiaceae): Comparative analysis and screening of high divergence sequences. *Sci Rep.* **2018**, *8*, 2210.
24. Rabah, S.O.; Shrestha, B.; Hajrah, N.H.; Sabir, M.J.; Alharby, H.F.; Sabir, M.J.; Alhebshi, A.M.; Sabir, J.S.; Gilbert, L.E.; Ruhlman, T.A. *Passiflora* plastome sequencing reveals widespread genomic rearrangements. *J Syst Evol.* **2018**, *56*, e12425.
25. Daniell, H.; Wurdack, K.J.; Kanagaraj, A.; Lee, S.-B.; Saski, C.; Jansen, R.K. The complete nucleotide sequence of the cassava (*Manihot esculenta*) chloroplast genome and the evolution of *atp*F in Malpighiales: RNA editing and multiple losses of a group II intron. *Theor Appl Genet.* **2008**, *116*, 723.
26. Asif, M.H.; Mantri, S.S.; Sharma, A.; Srivastava, A.; Trivedi, I.; Gupta, P.; Mohanty, C.S.; Sawant, S.V.; Tuli, R. Complete sequence and organisation of the *Jatropha curcas* (Euphorbiaceae) chloroplast genome. *Tree Genet Genomes* **2010**, *6*, 941-952.
27. Rivarola, M.; Foster, J.T.; Chan, A.P.; Williams, A.L.; Rice, D.W.; Liu, X.; Melake-Berhan, A.; Creasy, H.H.; Puiu, D.; Rosovitz, M. Castor bean organelle genome sequencing and worldwide genetic diversity analysis. *PloS one* **2011**, *6*, e21743.

28. Tangphatsornruang, S.; Uthaipaisanwong, P.; Sangsrakru, D.; Chanprasert, J.; Yoocha, T.; Jomchai, N.; Tragoonrung, S. Characterization of the complete chloroplast genome of *Hevea brasiliensis* reveals genome rearrangement, RNA editing sites and phylogenetic relationships. *Gene* **2011**, *475*, 104-112.
29. Li, Z.; Long, H.; Zhang, L.; Liu, Z.; Cao, H.; Shi, M.; Tan, X. The complete chloroplast genome sequence of tung tree (*Vernicia fordii*): Organization and phylogenetic relationships with other angiosperms. *Sci Rep.* **2017**, *7*, 1869.
30. Horvath, D.P.; Patel, S.; Doğramaci, M.; Chao, W.S.; Anderson, J.V.; Foley, M.E.; Scheffler, B.; Lazo, G.; Dorn, K.; Yan, C. Gene space and transcriptome assemblies of leafy spurge (*Euphorbia esula*) identify promoter sequences, repetitive elements, high-quality markers, and a full-length chloroplast genome. *Weed Sci.* **2018**, *66*, 355-367.
31. Shinozaki, K.; Ohme, M.; Tanaka, M.; Wakasugi, T.; Hayashida, N.; Matsubayashi, T.; Zaita, N.; Chunwongse, J.; Obokata, J.; Yamaguchi-Shinozaki, K. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *The EMBO journal* **1986**, *5*, 2043.
32. Kim, K.-J.; Lee, H.-L. Complete chloroplast genome sequences from korean ginseng (*Panax schinseng* nees) and comparative analysis of sequence evolution among 17 vascular plants. *DNA Res.* **2004**, *11*, 247-261.
33. Yi, D.-K.; Kim, K.-J. Complete chloroplast genome sequences of important oilseed crop *Sesamum indicum* L. *PLoS One* **2012**, *7*, e35872.
34. Powell, W.; Morgante, M.; McDevitt, R.; Vendramin, G.; Rafalski, J. Polymorphic simple sequence repeat regions in chloroplast genomes: Applications to the population genetics of pines. *Proc Natl Acad Sci U.S.A.* **1995**, *92*, 7759-7763.
35. Gupta, P.K.; Varshney, R. The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. *Euphytica* **2000**, *113*, 163-185.
36. Provan, J.; Powell, W.; Hollingsworth, P.M. Chloroplast microsatellites: New tools for studies in plant ecology and evolution. *Trends Ecol Evol.* **2001**, *16*, 142-147.
37. Timme, R.E.; Kuehl, J.V.; Boore, J.L.; Jansen, R.K. A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: Identification of divergent regions and categorization of shared repeats. *Am J Bot.* **2007**, *94*, 302-312.
38. Sherman-Broyles, S.; Bombarely, A.; Grimwood, J.; Schmutz, J.; Doyle, J. Complete plastome sequences from *Glycine syndetika*, and six additional perennial wild relatives of soybean. *G3-Genes Genomes Genet.* **2014**, g3. 114.012690.
39. Kawabe, A.; Nukii, H.; Furihata, H.Y. Exploring the history of chloroplast capture in *Arabis* using whole chloroplast genome sequencing. *Int J Mol Sci.* **2018**, *19*, 602.
40. Muriira, N.G.; Muchugi, A.; Yu, A.; Xu, J.; Liu, A. Genetic diversity analysis reveals genetic differentiation and strong population structure in *Calotropis* plants. *Sci Rep.* **2018**, *8*, 7832.
41. Saina, J.K.; Li, Z.-Z.; Gichira, A.W.; Liao, Y.-Y. The complete chloroplast genome sequence of tree of heaven (*Ailanthus altissima* (mill.)(Sapindales: Simaroubaceae), an important pantropical tree. *Int J Mol Sci.* **2018**, *19*, 929.
42. Kurtz, S.; Choudhuri, J.V.; Ohlebusch, E.; Schleiermacher, C.; Stoye, J.; Giegerich, R. Reputer: The manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* **2001**, *29*, 4633-4642.
43. Gu, C.; Dong, B.; Xu, L.; Tembrock, L.R.; Zheng, S.; Wu, Z. The complete chloroplast genome of *Heimia myrtifolia* and comparative analysis within Myrtales. *Molecules* **2018**, *23*, 846.
44. Li, Z.-Z.; Saina, J.K.; Gichira, A.W.; Kyalo, C.M.; Wang, Q.-F.; Chen, J.-M. Comparative genomics of the Balsaminaceae sister genera *Hydrocera triflora* and *Impatiens pinfanensis*. *Int J Mol Sci.* **2018**, *19*, 319.
45. Liu, H.-Y.; Yu, Y.; Deng, Y.-Q.; Li, J.; Huang, Z.-X.; Zhou, S.-D. The chloroplast genome of *Lilium henrici*: Genome structure and comparative analysis. *Molecules* **2018**, *23*, 1276.
46. Liu, X.; Li, Y.; Yang, H.; Zhou, B. Chloroplast genome of the folk medicine and vegetable plant *Talinum paniculatum* (Jacq.) Gaertn.: Gene organization, comparative and phylogenetic analysis. *Molecules* **2018**, *23*, 857.
47. Zhou, J.; Cui, Y.; Chen, X.; Li, Y.; Xu, Z.; Duan, B.; Li, Y.; Song, J.; Yao, H. Complete chloroplast genomes of *Papaver rhoeas* and *Papaver orientale*: Molecular structures, comparative analysis, and phylogenetic analysis. *Molecules* **2018**, *23*, 437.
48. Davis, C.C.; Latvis, M.; Nickrent, D.L.; Wurdack, K.J.; Baum, D.A. Floral gigantism in Rafflesiaceae. *Science* **2007**, *315*, 1812-1812.

49. Berry, P.E.; Hipp, A.L.; Wurdack, K.J.; Van Ee, B.; Riina, R. Molecular phylogenetics of the giant genus *Croton* and tribe Crotoneae (Euphorbiaceae sensu stricto) using its and *trn*L-*trn*F DNA sequence data. *Am J Bot.* **2005**, *92*, 1520-1534.
50. van Ee, B.W.; Berry, P.E. Taxonomy and phylogeny of Croton section *Heptallon* (Euphorbiaceae). *Syst Bot.* **2010**, *35*, 151-167.
51. van Ee, B.W.; Riina, R.; Berry, P.E. A revised infrageneric classification and molecular phylogeny of new world *Croton* (Euphorbiaceae). *Taxon* **2011**, *60*, 791-823.
52. Haber, E.A.; Kainulainen, K.; van Ee, B.W.; Oyserman, B.O.; Berry, P.E. Phylogenetic relationships of a major diversification of *Croton* (Euphorbiaceae) in the western indian ocean region. *Bot J Linn Soc.* **2017**, *183*, 532-544.
53. Doyle, J.J.; Doyle, J.L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull*. **1987**, *19*, 11-15.
54. Palmer, J.D. Isolation and structural analysis of chloroplast DNA. *Method Enzymol.* **1986**, *118*, 167-186.
55. Lowe, T.M.; Eddy, S.R. tRNAscan-se: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Aacids Res.* **1997**, *25*, 955-964.
56. Lohse, M.; Drechsel, O.; Kahlau, S.; Bock, R. OrganellarGenomeDraw—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Res*. **2013**, gkt289.
57. Darling, A.C.; Mau, B.; Blattner, F.R.; Perna, N.T. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **2004**, *14*, 1394-1403.
58. Edgar, R.C. Muscle: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. **2004**, *32*, 1792-1797.
59. Phobos – a tandem repeat search tool for complete genomes. http://www.rub.de/ecoevo/cm/cm_phobos.htm (archived on 5 May 2018)
60. Stamatakis, A.; Hoover, P.; Rougemont, J. A rapid bootstrap algorithm for the raxml web servers. *Syst Biol.* **2008**, *57*, 758-771.