Pertanika J. Trop. Agric. Sci. 44 (2): 429 - 447 (2021)

PERTANIKA JOURNALS

TROPICAL AGRICULTURAL SCIENCE

Journal homepage: http://www.pertanika.upm.edu.my/

In silico Characterization of the Structure of Genes and Proteins related to β-carotene Degradation in *Musa acuminata* 'DH-Pahang' and *Musa balbisiana* 'Pisang Klutuk Wulung'

Fenny Martha Dwivany[1,2,3]*, Nisrina Sukriandi[1], Karlia Meitha[1] and Tatas H. P. Brotosudarmo[4]

[1]School of Life Sciences and Technology, Institut Teknologi Bandung, Jl. Ganesa 10, Bandung 40132, West Java, Indonesia

[2]Bali International Research Center for Banana, Gedung Widyasaba lt. 3 Sayap Selatan Kampus UNUD Bukit Jimbaran Kuta Selatan Badung, Bali, Indonesia

[3]Bioscience and Biotechnology Research Center, Institut Teknologi Bandung, Jl. Ganesa 10, Bandung 40132, West Java, Indonesia

[4]Ma Chung Research Center for Photosynthetic Pigments (MRCPP) and Department of Chemistry, Universitas Ma Chung, Villa Puncak Tidar N01, Malang 65151, Indonesia

ABSTRACT

β-carotene is an important nutritional content in banana. However, its lifetime depends on the enzymes controlling its conversion into strigolactone. To understand the involved enzymes' activity, which are β-carotene isomerase (D27), carotenoid cleavage dioxygenase 7 (CCD7), and CCD8, would be the key to manipulate the rate of β-carotene degradation. In this research, the structure of genes and proteins of the D27, CCD7, and CCD8 from *Musa acuminata* were characterized. 'DH-Pahang' and *Musa balbisiana* 'Pisang Klutuk Wulung' (PKW). The corresponding sequence of genes from both species were aligned to determine similarity and intron/exon positions. Domains and motifs in the sequences of putative proteins of D27, CCD7, and CCD8 were also identified. It was found that *D27*, *CCD7*, and *CCD8* genes in DH-Pahang and PKW comprise of various nucleotide sequence length, putative proteins, and numbers and length of exons and introns. However, the putative proteins possess the same domains: DUF4033 (domain of unknown function) in D27 and RPE65 (retinal pigment epithelium) in CCD7 and CCD8. Phylogenetic trees showed that D27, CCD7, and CCD8 proteins from DH-Pahang and PKW are conserved and clustered in

ARTICLE INFO

Article history:
Received: 13 January 2021
Accepted: 26 March 2021
Published: 28 May 2021

DOI: https://doi.org/10.47836/pjtas.44.2.10

E-mail addresses:
fenny@sith.itb.ac.id (Fenny Martha Dwivany)
nisrina.sukriandi@gmail.com (Nisrina Sukriandi)
karliameitha@gmail.com (Karlia Meitha)
tatas.brotosudarmo@machung.ac.id (Tatas H. P. Brotosudarmo)
* Corresponding author

ISSN: 1511-3701
e-ISSN: 2231-8542

the same clades with the same proteins of monocot plants. Hence, the results could be useful for future research in optimizing β-carotene content in banana.

Keywords: A genome, B genome, β-carotene, *CCD*, *D27*

INTRODUCTION

Banana (*Musa* sp.) is a fruiting herb that is one of the most exported and consumed plants globally. Banana has many cultivars (Calberto et al., 2015). There are four genome types of banana, i.e. A, B, S, and T genomes. Banana cultivars mainly descended from *Musa acuminata* (A genome) and *Musa balbisiana* (B genome). The Cavendish cultivar of *M. acuminata* (AAA genome) dominates 90% of banana export value (Davey et al., 2013). The amount of β-carotene is 0.2 mg/100 grams (dry weight/d.w.) of fruit flesh, similar to *M. balbisiana* (0.19 mg/100 grams (d.w.) of fruit flesh) (Mathew & Muhammed, 2015). The β-carotene content in a plant is affected by three aspects: biosynthesis, degradation or conversion to another molecule, and storage in a plant. However, in other banana plants such as *Musa troglodytarum* (T genome) that contains a higher amount of β-carotene (1.5 mg/100 grams (d.w.) of fruit flesh) (Englberger et al., 2006), no significant differences were found when Buah (2015) compared the structure and expression of β-carotene biosynthesis genes between *M. troglodytarum* and *M. acuminata*. Therefore, we focus on the degradation aspect.

One of the carotenoid's degradation pathways is the conversion from carotenoid to strigolactone; a phytohormone that plays an essential role in inducing mutualistic symbiosis between fungi and plant roots in the form of mycorrhiza (Alder et al., 2012). Strigolactone plays a vital role in managing plant response to stress (Mishra et al., 2017). The pathway of β-carotene conversion to strigolactone involves three enzymes that are expressed from the genes *D27* (*DWARF27*) and *carotenoid cleavage dioxygenase 7* and *8* (*CCD7* and *CCD8*) (Alder et al., 2012).

Whole-genome sequencing of A genome for the cultivar DH-Pahang and B genome for the cultivar 'Pisang Klutuk Wulung' (PKW) has been conducted by D'Hont et al. (2012) and Davey et al. (2013). The genes in the whole genome sequence (WGS) of DH-Pahang have been predicted and available online at GenBank. However, the genes in the WGS of PKW have not been wholly annotated.

In this research, the structures of the genes *D27*, *CCD7*, and *CCD8* in *M. acuminata* cultivar DH-Pahang and *M. balbisiana* cultivar PKW were characterized *in silico*. In this work, the functions of the genes were also analyzed by identifying putative protein domains and motifs *in silico* and constructing phylogenetic trees for putative protein sequences of D27, CCD7, and CCD8. From this study, researchers were able to understand more about the genes, and hope that it could be beneficial for future biofortification efforts in banana.

MATERIALS AND METHODS

WGS of DH-Pahang and PKW Sequence Data Retrieval

D27, *CCD7*, and *CCD8* genes in the WGS of DH-Pahang were annotated *in silico* and available in NCBI GenBank (Clark et al., 2016). The gene sequences were predicted *in silico* from the annotation of genes in Eukaryotes (in this case, the *M. acuminata*). Meanwhile, the genetic sequences of PKW were retrieved by BLAST (Basic Local Alignment Search Tool) (Altschul et al., 1990) from the genes of DH-Pahang to the WGS of PKW, by using the application blast2seq in NCBI with the algorithm optimization megablast so that the resulting sequences are highly similar to the genes from DH-Pahang. Sequences with the highest max score were chosen and then predicted to be the nucleotide sequences of *D27*, *CCD7*, and *CCD8* in PKW. For reference, a search for those genes in banana genomes based on the same genes in *Arabidopsis thaliana* as a model organism was also done.

Prediction of Gene Structure, Motifs, and Domains of D27, CCD7, and CCD8 Putative Proteins of DH-Pahang and PKW

Prospective nucleotide sequences were then predicted and annotated to obtain gene structure, putative amino acid sequences, and predicted protein motifs and domains. To predict gene structure and putative amino acid sequences, the program FGENESH (Solovyev et al., 2006) was used. FGENESH predicts exon-intron structure in a gene and predicts putative amino acid sequences. Exon and intron mapping were also validated by sim4 (Florea et al., 1998), by mapping mRNA putative sequences into respective genes' sequences. Exon and intron structures of PKW's genes were then compared to DH-Pahang's genes. Then, protein motifs and domains were predicted using the NCBI Conserved Domain Search (Marchler-Bauer et al., 2017) and MEME-Suite (Bailey & Elkan, 1994). Motif identification was done using the motif search in InterProScan (Jones et al., 2014).

Comparison of Gene Structure and Putative Proteins of D27, CCD7, and CCD8 between DH-Pahang and PKW

Predicted nucleotide structure of *D27*, *CCD7*, and *CCD8* were compared for each gene, between DH-Pahang and PKW. Gene structures compared were nucleotide sequence length, number of exons and introns, and position for each exon and intron. Predicted protein structures were compared between DH-Pahang and PKW; the comparisons were motif and domain positions in the protein and length of amino acid sequences that made up those motifs and domains. Aside from that, pairwise alignment between the respective genes and proteins between the two cultivars was determined to determine similarity and identity percentage between the genes/proteins. The pairwise alignment was done by using EMBOSS-Needle (Madeira et al., 2019).

Construction of Phylogenetic Trees of D27, CCD7, and CCD8 Genes and Putative Proteins between DH-Pahang, PKW, and Other Plant Species

Phylogenetic analyses were done to determine the similarity and genetic relation of D27, CCD7, and CCD8 putative proteins in A and B genome bananas and other monocot plants with similar genes (paralogous to banana). The phylogenetic tree was made based on the putative amino acid sequence of D27, CCD7, and CCD8 in DH-Pahang, PKW, and other plants. Amino acid sequences of those proteins in different plants were obtained by using BLAST towards the GenBank database. Aside from paralogous sequences, nucleotide and amino acid sequences from *A. thaliana* were also retrieved to serve as an outgroup in the phylogenetic tree.

After sequences were retrieved, they were aligned globally with T-Coffee v.11, UNIX-based (Notredame et al., 2000), and trimmed with BioEdit v.7.2.6 (Hall, 1999). Phylogenetic trees were constructed with MrBayes (Huelsenbeck & Ronquist, 2001; Ronquist & Huelsenbeck, 2003). Parameters used were invgamma for rates (among-site rate variation), nGen (number of generations) variable from 10,000 to 20,000 trees, with sampling frequency for every 100 trees. Phylogenetic trees constructed were then visualized with TreeView v.1.6.6 (Page, 1996).

RESULTS AND DISCUSSION

D27, *CCD7*, and *CCD8* Genes in DH-Pahang and PKW Based on Reference Nucleotide Sequences in GenBank and Genetic Sequences of *Arabidopsis thaliana*

Table 1 shows a summary of the genes (detailed in Figure 1) DH-Pahang and PKW. Figures 2-4 show the exon-intron structure (part-a), the protein domain structure (part-b), and the putative protein motive structure (part-c) in genes *D27*, *CCD7* as well as *CCD8* in DH-Pahang and PKW.

Table 1
*Genes involved in the conversion of β-carotene into strigolactone in DH-Pahang (*Musa acuminata*) and PKW (*Musa balbisiana*)*

No.	Gene name	Species	GeneID	Seq. length (bp)	Located in chromosome no.	Number of exons	% Similarity
1	*D27*	*Musa acuminata*	103986218	3,532	1	7	41
		Musa balbisiana		2,602		5	
		Musa acuminata	103976367	2,891	2	6	97.2
		Musa balbisiana		2,897		5	
2	*CCD7*	*Musa acuminata*	103972006	3,768	11	7	95.4
		Musa balbisiana		3,723		5	
3	*CCD8*	*Musa acuminata*	103975947	2,651	2	6	40.4
		Musa balbisiana		2,658		5	
		Musa acuminata	103989799	2,790	6	6	93.5
		Musa balbisiana		2,685		6	

Note. % Similarity value was obtained through global pairwise alignment with EMBOSS-Needle

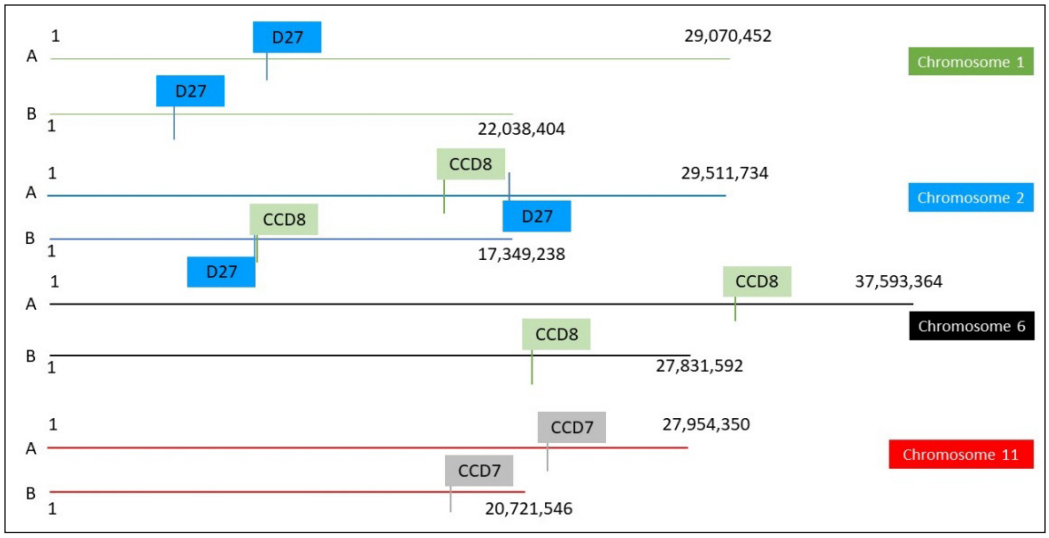

Figure 1. Positions of *D27*, *CCD7*, dan *CCD8* genes at chromosomes of DH-Pahang (labelled A) and PKW (labelled B)

Note. Each chromosome is given different colours, and its sequence length is labelled. Each gene is labelled with different colours: *D27* in blue, *CCD7* in grey, and *CCD8* in green

The search and prediction of genes based on nucleotide sequences in GenBank and genetic sequences of *A. thaliana* results in identical or similar nucleotide sequences and have similar positions in both DH-Pahang's genome and PKW's. However, there are differences such as the prediction results of the gene *D27* based on *A. thaliana* can only be found on chromosome 1; meanwhile, according to the prediction results based on the reference sequence from GenBank, the *D27* gene was found on chromosomes 1 and 2.

Genetic Structure Comparison of *D27* Gene in DH-Pahang and PKW

Figure 2 shows the exon-intron structure (part-a), the protein domain structure (part-b), and the putative protein motif structure (part-c) of *D27* in DH-Pahang and PKW (motif details available in Table S1).

D27 genes in DH-Pahang were retrieved from GenBank. *D27* was found in chromosomes 1 and 2, with GeneID 103986218 for *D27* at chromosome 1 and 103976367 for *D27* at chromosome 2. The nucleotide sequence length of the *D27* gene at chromosome 1 of DH-Pahang was 3,532 base pairs (bp) with seven exons, and at chromosome 2 was 2,891 bp with five exons. *D27* could also be found at chromosome 1 and 2 of PKW's whole genome sequence. The nucleotide sequence length of the *D27* gene at chromosome 1 was 2,891 bp and at chromosome 2 was 2,897 bp. In *A. thaliana*, the number of exons in the *D27*-like gene (GeneID: 838334) was seven (Waters et al., 2012). Sequence alignment between the *D27* gene in chromosome 1 of DH-Pahang and PKW resulted in a similarity percentage between the two sequences of 41%. *D27* gene in chromosome 2 of DH-Pahang and

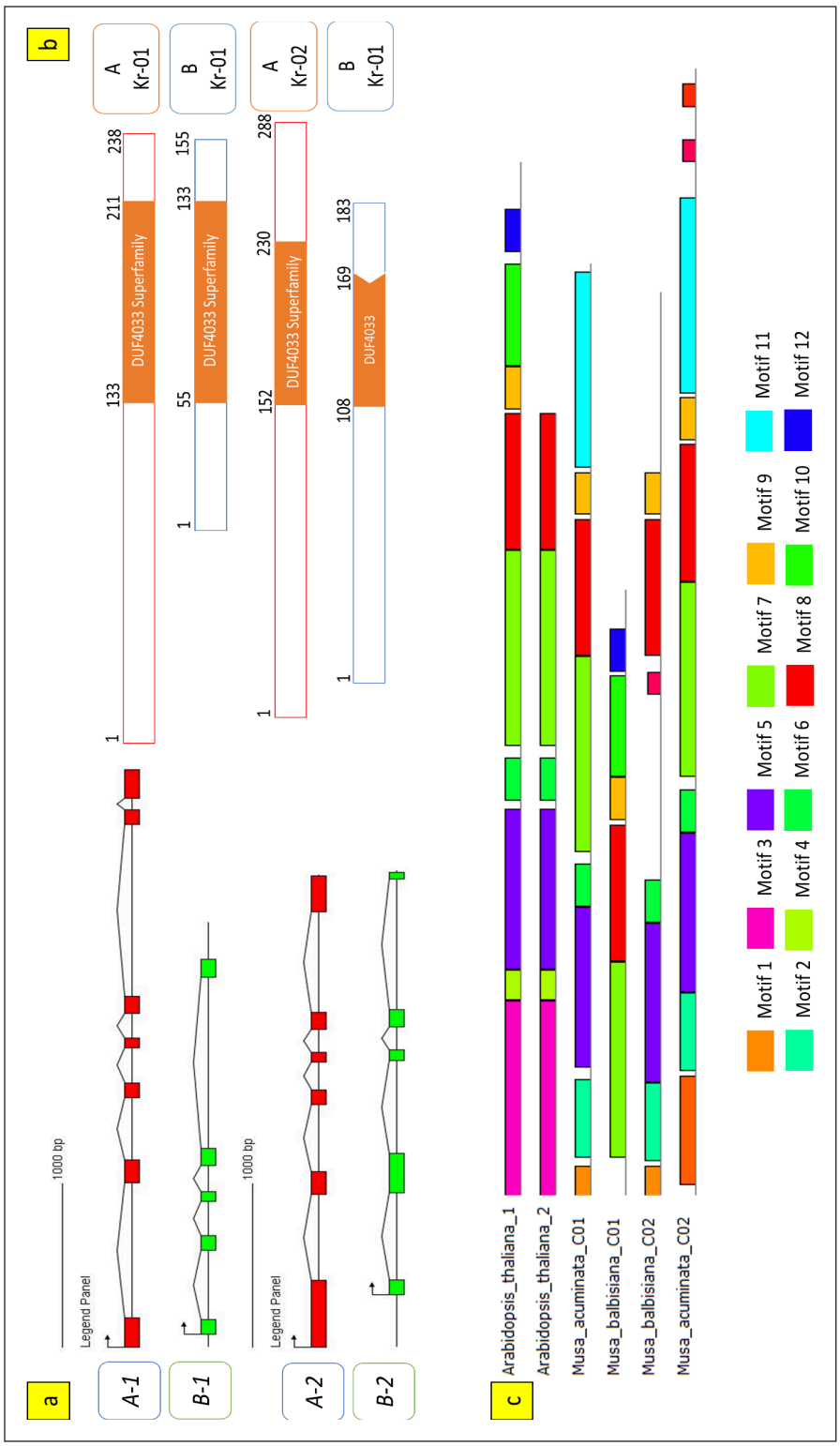

Figure 2. Comparison of exon-intron structure, protein domain and motif structure in *D27*
Note. a) Comparison of exon-intron structures of *D27* gene in chromosome 1 and 2 DH-Pahang (*Musa acuminata*) (A-1, A-2) and chromosome 2 PKW (*Musa balbisiana*) (B-1, B-2); b) Domain structure of D27 putative protein as the product of the gene in chromosome 1 and 2 in DH-Pahang (A) and PKW (B). Kr-01: chromosome 1 and Kr-02: chromosome 2. Numbers at the side indicate the number of amino acid residue in a sequence; **c)** Motifs of D27 putative protein sequences in DH-Pahang, PKW, and *Arabidopsis thaliana* predicted with MEME-Suite

Table S1
Motif consensus sequence of D27 putative protein and result of motif search with InterProScan

D27 Motif	E-value	Length (aa)	Available at the database?	Motif consensus sequence
1	2.4e+003	8	No	MELGQQRP
2	3.4e-004	20	No	RKLLSSVVEARVQTEEKMVA
3	1.7e-016	50	No	MNTKLSLSQTKIFTFTTWFNDTRSGLDRRSSISPTLCSK-PVYSGKLKAAK
4	3.7e+003	8	No	ETARIETS
5	2.7e-325	41	Yes**	ATEKTVYKDNWFDKLAIGYLSRNLQEASGMKNEKDGYESLI
6	5.7e-038	11	No	EAAJMISRLFD
7	2.0e-406	50	Yes**	QQELVIQALERAFPSYILTMIKVMLPPSKFSREYFAAFT-TIFFPWLVGPC
8	6.9e-395	35	Yes*	EVRESEVDGRKEKNVVYIPKCRFLESTNCVGMCTN
9	3.4e-104	11	No	CKIPSQKFIQD
10	6.8e-003	26	Yes***	SLGMPIYMEPDFEDLSCEMIFGRZPP
11	1.9e-424	50	Yes*	LGMPVYMSPNFEDMSCEMIFGQQPPEDDPALKQPCYRTK-CIAKQNHGVNC
12	4.0e+000	11	No	DDPALKQPCYH

Note. *) Motif: Beta-carotene isomerase D27-like (IPR038938); in domain of unknown function DUF4033 (IPR025114); **) Motif: Beta-carotene isomerase D27-like (IPR038938); ***) Motif in domain of unknown function DUF4033 (IPR025114)

PKW was 97.2% similar; hence it could be predicted that *D27* in chromosome 2 as more conserved between two different *Musa* species.

Domain and Motif of D27 Putative Protein and Comparison between DH-Pahang and PKW

D27 putative protein's amino acid sequences in DH-Pahang were retrieved from GenBank (accession number XP_009402431 for D27 in chromosome 1 and XP_018677614 for D27 in chromosome 2).

D27 putative protein's amino acid sequences in PKW were predicted through their respective nucleotide sequences using FGENESH. All D27 proteins have the domain DUF4033, and they are in the protein family DUF4033. DUF4033 (domain of unknown function 4033) is a domain with a function that has not been characterized. This domain can be found in bacteria and eukaryotes, 80 amino acids-long (Marchler-Bauer et al., 2017). Most DUFs are highly conserved, and this suggests their essential role in biological function. However, most DUFs are non-essential, so their biological role became difficult to be determined. DUFs are believed to be only needed under certain conditions (Häuser et al., 2012).

Based on motif prediction with MEME-Suite of 16 D27 protein sequences from 13 plant species, protein motifs found are shown in Figure 2c. The motif sequence consensus was then used as a query searching for the types of motifs and protein families

containing these motifs using InterProScan. Five motifs were found in the database. The existence of unidentified motifs could be estimated as novel motifs that need to be studied further regarding their role in this protein's function. Based on the results of the search for motifs in InterProScan, the five motifs found in the database were those found in the D27-like β-carotene isomerase protein family (IPR0389038) and were incorporated in the DUF4033 domain component.

The motif structure of each D27 protein sequence from DH-Pahang, PKW, and *A. thaliana* are as shown in Figure 2c. All D27 proteins in DH-Pahang, PKW, and *A. thaliana* had motifs that form the B-carotene isomerase domain (motifs 5, 7, 8, and 10). Motifs 1 and 2 were only found in DH-Pahang and PKW, while motifs 3 and 4 were only found in *A. thaliana*. Motif 11 appeared only on *M. acuminata*. In general, the genus *Musa* has motifs 1, 2, 5, 6, 7, 8, and 9. However, in the putative protein sequence D27 *M. balbisiana* from chromosome 1, the only motifs shared by other *Musa* species were motifs 7, 8, and 9. This could be because the putative protein sequences were cut off at the front and back because the protein sequence prediction using FGENESH was not perfect.

CCD7 Genetic Structure Comparison between DH-Pahang and PKW

Figure 3 shows the exon-intron structure (part a), the protein domain structure (part b), and the putative protein motif structure (part c) in *CCD7* gene in DH-Pahang and PKW. Detailed motif consensus sequence of CCD7 putative protein and result of motif search are available in Table S2.

The *CCD7* gene in DH-Pahang was obtained from GenBank. This gene was found on chromosome 11, with GeneID 103972006. The *CCD7* gene on chromosome 11 DH-Pahang had a sequence length of 3,768 bp. The *CCD7* gene BLAST results from the PKW whole genome sequence were found on chromosome 11, with a sequence length of 2,897 bp. The similarity of *CCD7* genes in *M. acuminata* and *M. balbisiana* was 95.4%. This indicates that this gene was sufficiently conserved in two plant species in the same genus. This was in accordance with the study by Ahrazem et al. (2016): that *CCD7* is a well-conserved gene between plant species; this gene has a function crucial to the organism's survival, hence its structure is maintained. The number of exons in *CCD7* DH-Pahang was 7, while in PKW, there were 6. This was in accordance with the results of research by Wang et al. (2017), the number of *CCD7* exons of various plants ranged from 5 to 7.

Domain and Motif of CCD7 Putative Protein and Comparison between DH-Pahang and PKW

DH-Pahang's CCD7 putative protein sequence was retrieved from GenBank (accession number XP_009384463). The putative protein sequence of the *CCD7* gene on PKW was obtained from gene prediction results in FGENESH. CCD7 proteins belong to the RPE65 domain, and all of them belong to the RPE65 protein superfamily. The RPE65 domain (retinal pigment epithelium

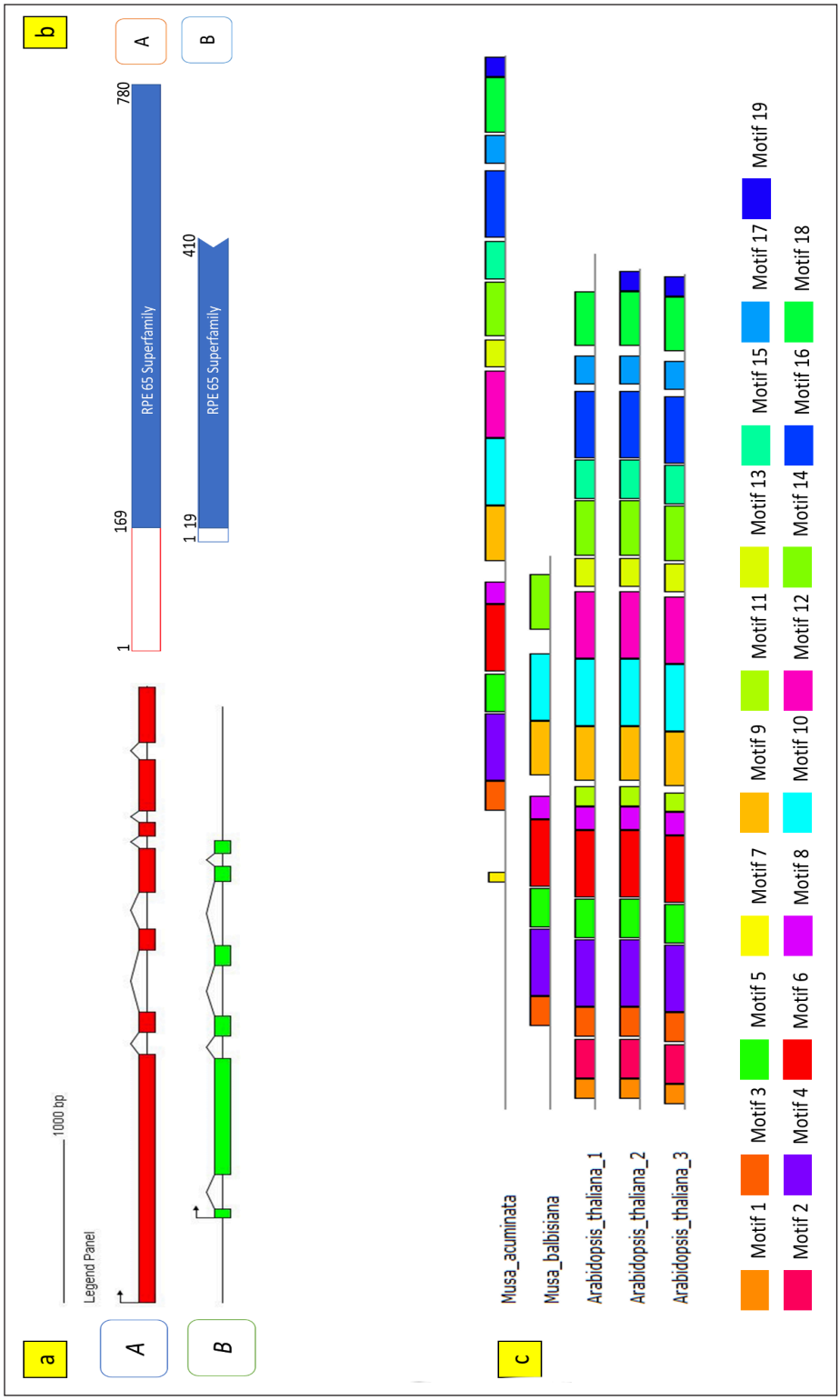

Figure 3. Comparison of exon-intron structure, protein domain, and motif structure in *CCD7*
Note. a) Comparison of exon-intron structures of *CCD7* gene in chromosome 11 DH-Pahang (*Musa acuminata*) (A) and PKW (*Musa balbisiana*) (B); **b)** Domain structure of CCD7 putative protein as the product of the gene in DH-Pahang (A) and PKW (B). Numbers at the side indicate the number of amino acid residue in a sequence; **c)** Motifs of CCD7 putative protein sequences in DH-Pahang, PKW, and *Arabidopsis thaliana* predicted with MEME-Suite

Table S2
Motif consensus at CCD7 putative protein and search result with InterProScan

CCD7 Motif	E-value	Length (aa)	Available at the database?	Motif consensus sequence
1	2.5e-034	15	No	IPPKLLPPAKLPPTH
2	3.9e-052	29	No	HGQTNLPLAESKCKDSWSMPDDNMVRLGT
3	3.7e-186	22	No	PDSTSAAFWDYQFLFVSQRSET
4	1.1e-515	50	Yes*	AEPVVLRVVEGSIPVDFPSGTYYLAGPGLFTDDHGSTVH-PLDGHGYLRAF
5	1.3e-176	29	No	IDGSSGQVKFSARYVETEAQREERDPVTG
6	5.2e-620	50	Yes*	WRFTHRGPFSVLKGGKRVGNTKVMKNVANTSVLRWGGRLL-CLWEGGDPY
7	2.7e-018	8	-	MQAKPCHN
8	1.5e-080	17	No	IDSRTLDTVGKFDLIGN
9	2.8e-375	41	No	FLDVAAHLLKPILYGVFKMPPKRLLSHYKIDARRNRLLMVS
10	1.3e-615	50	No	CNAEDMLLPRSNFTFYEFDSNFELLQKKEFVIPDHLMIH-DWAFTDTHYIL
11	4.1e-011	15	No	GCESCDDDDSSDRDL
12	6.4e-367	50	No	FGNRIKLDIPGSLLAVCGLSPMISALSVNPSKPSSPI-YLLPRFSDKEARG
13	5.1e-185	21	No	WRVPIEAPSQLWVLHVGNAFE
14	5.8e-466	41	No	DENGNLNIQJQASGCSYQWFNFQKMFGYBWQSGKLDPSFMN
15	5.0e-181	29	No	EEGEEKLLPHLVQVSINLDSTGNCTRCSV
16	3.2e-393	50	Yes*	LSNQWNKPADFPAINPDFSGRKNKYVYAATSSGSRR-FLPHFPFDSVVKLB
17	1.6e-145	21	No	VRTWSAGARRFIGEPVFVPRG
18	6.2e-369	41	Yes*	EDDGYILVVEYAVSTQRCYLVILDAKKIGEKBAVVARLEVP
19	2.1e-113	15	No	KHLTFPLGFHGFWAD

Note. *) Carotenoid oxygenase protein family (IPR004294); domain RPE65 (PF03055). Sequence written in red font was not checked at InterProScan because its length is less than ten amino acids

protein) is a domain belonging to a protein family widely expressed in the retinal pigment epithelium. This protein family also consists of enzymes that can cut neoxanthin in plants, and lignostilbene-alpha, β-dioxygenase enzymes in bacteria. The CCD protein family is characterized by the presence of the RPE65 domain (Marchler-Bauer et al., 2017). Neoxanthin is a form of a carotenoid compound.

Based on motif prediction of 15 CCD7 amino acid sequences from 13 plant species, protein motifs found are shown in Figure 3c. The motif sequence consensus is then used as a query searching for the types of motives and protein families containing these motifs using InterProScan. Four motifs were found in the database. The existence of motifs which functions have not been identified could be estimated as novel motifs

that need to be studied further regarding their role in this protein's function. Based on the results of the search for motives in InterProScan, the four motifs found in the database were those found in the carotenoid oxygenase protein family (IPR004294) and incorporated in the RPE65 domain component.

Meanwhile, protein prediction results concluded that proteins in the family carotenoid oxygenase are proteins involved in the reduction-oxidation process (GO: 0055114) and have a specific molecular function: oxidoreductase activity (GO: 0016702). The motif for each CCD7 amino acid sequence in DH-Pahang, PKW, and *A. thaliana* is shown in Figure 3c. All CCD7 putative proteins have the carotenoid dioxygenase motif (motifs 4, 6, 16, and 17). Motifs 1 and 2 were only found at *A. thaliana*, and motif 7 was only found at *M. acuminata*. Both *M. acuminata* and *M. balbisiana* have the motifs 3, 4, 5, 6, 8, 9, 10, and 11. However, in *M. acuminata,* some motifs were not found at the CCD7 amino acid sequence of *M. balbisiana*. This could be because the putative protein sequences were cut off at the front and back because the protein sequence prediction using FGENESH was not perfect.

CCD8 Genetic Structure Comparison between DH-Pahang and PKW

Figure 4 shows the exon-intron structure (part a), the protein domain structure (part b), and the putative protein motif structure (part c) of *CCD8* in DH-Pahang and PKW. Detailed motif consensus sequence of CCD8 putative protein and result of motif search are available in Table S3.

The *CCD8* gene in DH-Pahang was retrieved from GenBank. This gene was found on chromosomes 2 and 6, with GeneID 103975947 for *CCD8* on chromosome 2 and 103989799 for *CCD8* on chromosome 6. The *CCD8* gene on chromosome 2 DH-Pahang had the length of 2,651 base pairs, and on PKW 2,123 base pairs. The *CCD8* gene from BLAST resulted from the PKW whole genome sequence was also found on two chromosomes: chromosome 2 and chromosome 6 from PKW. The number of exons in the *CCD8* gene on chromosome 2 DH-Pahang was 6, while in PKW, there were 5. Sequence alignment analysis showed the similarity between the *CCD8* gene on chromosome 2 DH-Pahang and PKW was 40.5%. The *CCD8* gene on chromosome 6 DH-Pahang had 2,790 base pairs, while the *CCD8* gene from chromosome 6 had 2,685 base pairs. The number of exons in the *CCD8* gene from chromosome 6 DH-Pahang and PKW was five. After sequence alignment between the *CCD8* gene on chromosome 6 DH-Pahang and PKW, the percentage values of identity and similarity between these two sequences were 93.5%. The number of exons in the *CCD8* genes in DH-Pahang and PKW was five to six; consistent with the results of a study by Batra et al. (2019) regarding *CCD8* genes and proteins in seven monocot and dicot species: the number of exons in the *CCD8* gene generally ranges from four to seven.

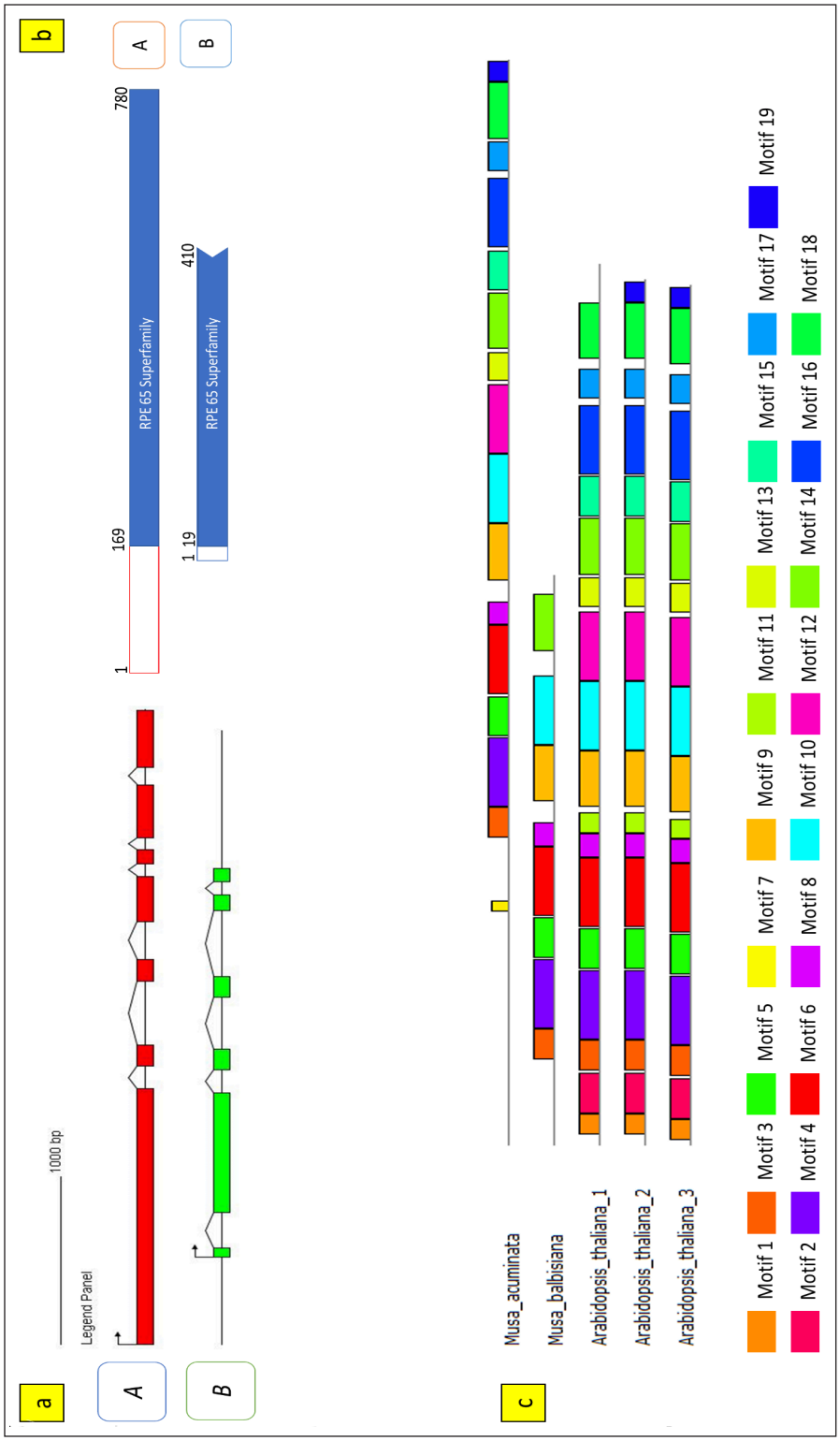

Figure 4. Comparison of exon-intron structure, protein domain, and motif structure in *CCD8*

Note. a) Comparison of exon-intron structures of *CCD8* gene in chromosome 2 and 6 in DH-Pahang (*Musa acuminata*) (A-2, A-6) and chromosome 2 and 6 in PKW (*Musa balbisiana*) (B-2, B-6); b) Domain structure of CCD8 putative protein as the product of the gene in DH-Pahang (A) and PKW (B). Kr-02: chromosome 2 and Kr-06: chromosome 6. Numbers at the side indicate the number of amino acid residue in a sequence; **c)** Motifs of CCD8 putative protein sequences in DH-Pahang, PKW, and *Arabidopsis thaliana* predicted with MEME-Suite

Table 3
Motif consensus sequence of CCD8 putative protein and result of motif search with InterProScan

CCD8 Motif	E-value	Length (aa)	Available at the database?	Motif consensus sequence
1	1.9e-028	50	No	TQFSSPKAHASHAHVAVSTRPGSVYSGNSIGDAVNK-SKPHVPGGLRARRV
2	5.1e-005	21	No	ETQVAPEPQPEPEKGGGEERK
3	6.1e-020	21	No	RRPAESVRASVATEPRPTVPS
4	2.6e+000	15	No	FDPAVETKQDVGSGR
5	1.9e-002	20	No	MASTLFSPPLQPTAIFPSTR
6	2.1e+003	7	-	FDTICRR
7	1.1e-588	50	Yes*	AWTSIRQERWEGELVVZGEIPLWLNGTYLRNGPGL-WNIGDYNFRHLFDGY
8	3.0e-523	50	Yes**	TLVRLHFENGRLIAGHRQIESEAYKAAKKNKRL-CYREFSEVPKPDNFLAY
9	1.3e-012	6	-	GELASL
10	7.4e-588	50	Yes*	FSGASLTDNANTGVVKLGDGRVVCLTETIKGSIQID-PDTLETIGKFEYSD
11	1.9e-569	50	Yes**	GLIHSAHPIVTESEFLTLLPDLVRPGYLVVRMEPGS-NERKVIGRVNCRG
12	3.4e-666	50	Yes*	WVHSFPVTEHYVVVPEMPLRYCAQNLLRAEPTPLYK-FEWHPESGSFMHVM
13	1.5e-331	29	No	CKASGKIVASVEVPPYVTFHFINAYEEKD
14	1.2e-580	50	Yes**	TAIIADCCEHNADTTILDKLRLQNLRSFSGEDVLP-DARVGRFRIPLDGSP
15	1.3e-060	11	No	GELEAALDPEE
16	1.8e-677	50	Yes*	HGRGMDMCSINPAYLGKKYRYAYACGAQRPCNF-PNTLTKIDLVEKKAKNW
17	5.9e-094	11	No	GAVPSEPFFVA
18	1.9e-589	50	Yes*	RPGATEEDDGVVISMVSDKNGEGYALLLDGSTFEEI-ARAKFPYGLPYGLH
19	8.1e-050	6	-	GCWVPK

Note. *) Carotenoid oxygenase protein family (IPR004294) and domain RPE65 (PF03055); **) Carotenoid oxygenase protein family (IPR004294). Sequences written in red font were not checked at InterProScan because their lengths are less than ten amino acids

Domain and Motif of CCD8 Putative Protein and Comparison between DH-Pahang and PKW

The amino acid sequences for the putative protein CCD8 from DH-Pahang were obtained from GenBank (accession number XP_009389368 for chromosome 2 and XP_009407022 for chromosome 6). The amino acid sequences for the putative protein of the CCD8 gene on chromosomes 2 and 6 PKW were obtained from gene prediction results in FGENESH. All CCD8 proteins belong to the RPE65 domain, and like other CCD proteins, they all belong to the RPE65 protein superfamily (Marchler-Bauer et al., 2017). The RPE65 domain

(retinal pigment epithelium protein) is the domain that belongs to the entire CCD protein family. This domain is characteristic of enzymes involved in apocarotenoid biosynthesis, the intermediate product of the synthesis of strigolactone from β-carotene (Batra et al., 2019).

Based on the prediction of protein motifs of 16 CCD8 protein sequences from 13 plant species, the protein motifs found are shown in Figure 4c. The motif sequence consensus was then used as a query in searching for the types of motifs and protein families containing these motifs using InterProScan. Four motifs were found in the database. The existence of motifs which functions had not been identified could be predicted as novel motifs that needed to be studied further regarding their role in this protein's function. Based on the motif search results in InterProScan, the eight motifs found in the database were those found in the carotenoid oxygenase protein family (IPR004294) and incorporated in the RPE65 domain component.

Meanwhile, protein prediction results concluded that proteins in the carotenoid oxygenase family were proteins involved in the reduction-oxidation process (GO: 0055114). They have a particular molecular function of oxidoreductase activity (GO: 0016702). This corresponds to the putative protein sequence CCD7, a protein in the same family as the CCD8 protein (Wang et al., 2017). The motif structures of each CCD8 protein sequence from DH-Pahang, PKW, and *Arabidopsis thaliana* are shown in Figure 4c. All CCD8 putative proteins had a carotenoid oxygenase motif (motifs 7, 8, 10, 11, 12, 14, and 16). There were specific motifs only in monocot or dicot plant species: motifs 1 and 2 were only found in *A. thaliana*. In contrast, motifs 3 and 4 were only found in putative protein products CCD8 from chromosome 6 DH-Pahang and PKW. Motif 5 was found only in *M. acuminata*. In general, CCD8 putative protein sequences in the genus *Musa* had motifs 3, 5, 7, 8, 9 to 19. However, in the putative protein sequences of CCD8 chromosome 2 *M. balbisiana*, there were no motifs before motif 7, and this could be because the putative protein sequences were cut off at the front and back because the protein sequence prediction using FGENESH was not yet perfect. The difference between the motifs in *M. acuminata* and *M. balbisiana* could also be caused by the mutation process of insertion and deletion, along with the evolutionary process of the two species.

Phylogenetic Relations of Amino Acid Sequences of D27, CCD7, and CCD8 Putative Proteins in DH-Pahang, PKW, and Other Plant Species

A phylogenetic tree was built from putative protein D27 on chromosome 1 and 2 DH-Pahang and putative protein D27 from PKW and other plants; as shown in Figure 5.

Figure 5 shows that the putative protein D27 from chromosome 1 DH-Pahang (*M. acuminata*) was present in the same clade as putative protein D27 from chromosome 1 PKW (*M. balbisiana*); the same goes for putative protein D27 from chromosome 2. Putative protein D27 from the genus *Musa*

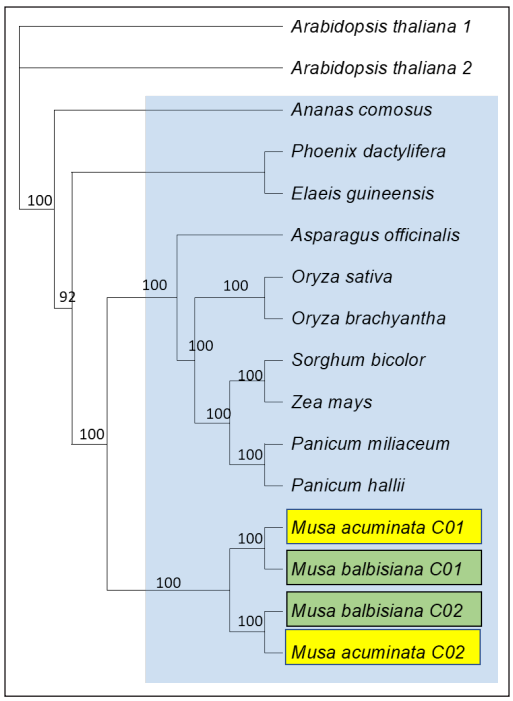

Figure 5. Phylogenetic tree based on putative protein sequences of D27 DH-Pahang (*Musa acuminata*), PKW (*Musa balbisiana*), and D27 protein from other plants

Note. C01 = chromosome 1, C02 = chromosome 2. The blue square indicates monocot plants. The tree was constructed using the Bayesian algorithm with MrBayes software, nGen = 20,000, and mutation rate = invgamma

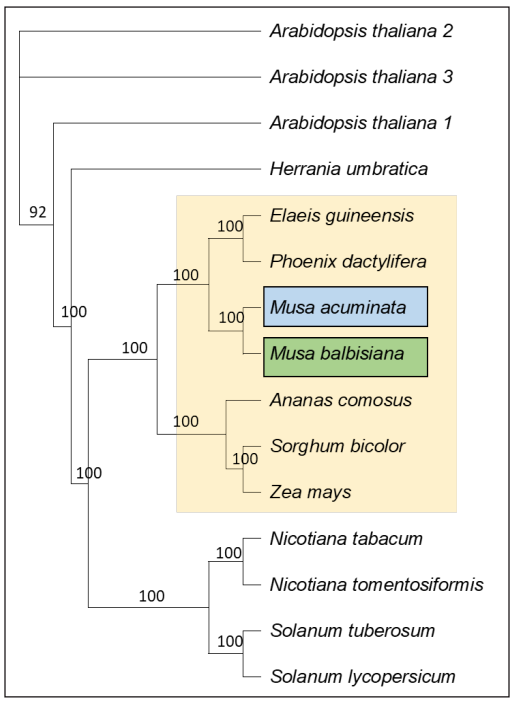

Figure 6. Phylogenetic tree based on putative protein sequences CCD7 DH-Pahang (*Musa acuminata*), PKW (*Musa balbisiana*), and CCD7 protein from other plants

Note. The yellow square indicates monocot plants. The tree was constructed using the Bayesian algorithm with MrBayes software, nGen = 20,000, and mutation rate = invgamma

was found in a separate clade from other species, but still had a common ancestor with monocot species such as species in the genus *Panicum* and *Oryza*, as well as species such as *Sorghum bicolor* and *Asparagus officinalis*.

The phylogenetic tree based on the putative protein sequence CCD7 in DH-Pahang, PKW, and other plants is shown in Figure 6. The CCD7 putative protein in DH-Pahang (*M. acuminata*) was present in the same clade as the putative protein CCD7 PKW (*M. balbisiana*). The *Musa* clade has a common ancestor with the clade *Elaeis guineensis* - *Phoenix dactylifera*. This clade was a sister group of clades of monocot plants such as *Ananas comosus*, *Sorghum bicolor*, and *Zea mays*. Meanwhile, other species were dicot plants; for example, the genera *Nicotiana* and *Solanum* belonging to the Solanaceae family and were separated from the clades of monocot plants.

The phylogenetic tree based on the putative protein sequence CCD8 from chromosomes 2 and 6 in DH-Pahang, PKW, and other plants is shown in Figure 7.

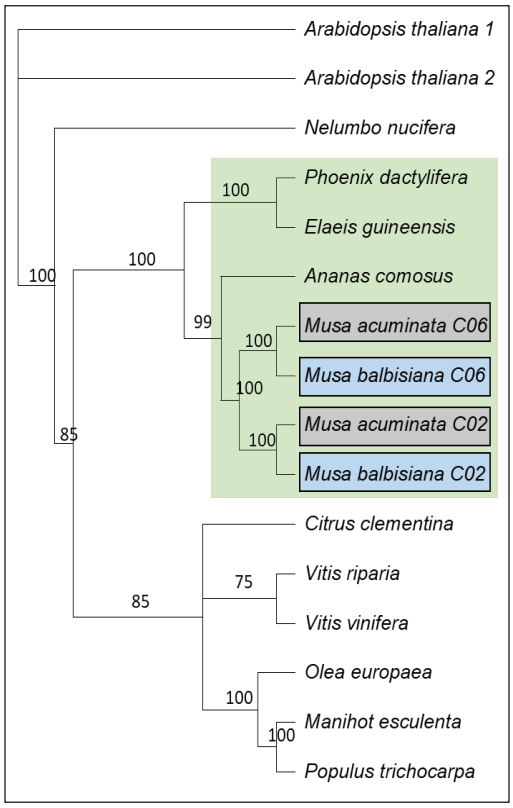

Figure 7. Phylogenetic tree based on putative protein sequences CCD8 DH-Pahang (*Musa acuminata*), PKW (*Musa balbisiana*), as well as CCD8 protein from other plants

Note. C02 = chromosome 2, C06 = chromosome 6. The green square indicates monocot plants. The tree was constructed using the Bayesian algorithm with MrBayes software, nGen = 20,000, and mutation rate = invgamma

The putative protein CCD8 from chromosome 2 DH-Pahang (*M. acuminata*) was in the same clade as the respective protein in PKW (*M. balbisiana*); likewise, putative protein CCD8 from chromosome 6. The putative protein CCD8 from the genus *Musa* was in the same clade as *Ananas comosus*, a monocot plant. These three species also belonged to the same clade as *Elaeis guineensis* and *Phoenix dactylifera*, both monocot plants. This monocot plant group was separate from other plants, which included the dicot group.

Based on this phylogenetic analysis results, it can be concluded that all putative protein sequences of DH-Pahang and PKW were the most similar to monocot plant species.

The putative protein sequences D27, CCD7, and CCD8 of the monocot species were in separate clades with dicot species. This suggested that the *D27*, *CCD7*, and *CCD8* genes diverged at an early stage in monocots and dicots' evolutionary history. This was consistent with research conducted by Batra et al. (2019), in which they also found that amino acid sequences of the *CCD8* gene of seven monocot species and eight dicot species showed a divergence between monocot-dicot.

CONCLUSION

This study concluded that the *D27*, *CCD7*, and *CCD8* genes in genomes A and B have different nucleotide sequence lengths and putative protein and different numbers and positions of exon-introns. Then, putative protein D27 has the DUF403 domain, whereas putative protein CCD7 and CCD8 have the RPE65 domain as the marker domain of each family.

The next steps are to isolate and sequence the *D27*, *CCD7*, and *CCD8* genes in DH-Pahang and PKW; and characterize their protein products in further research. In addition, there were still unidentified putative

protein motifs, and they may be essential motifs as markers and differentiators of the D27, CCD7, and CCD8 proteins in *Musa* compared to other plant species.

ACKNOWLEDGEMENTS

This study was funded by Institut Teknologi Bandung Priority Research KK A 2020 for Fenny M. Dwivany.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

Ahrazem, O., Gómez-Gómez, L., Rodrigo, M. J., Avalos, J., & Limón, M. C. (2016). Carotenoid cleavage oxygenases from microbes and photosynthetic organisms: Features and functions. *International Journal of Molecular Sciences*, *17*(11), 1781. https://doi.org/10.3390/ijms17111781

Alder, A., Jamil, M., Marzorati, M., Bruno, M., Vermathen, M., Bigler, P., Ghisla, S., Bouwmeester, H., Beyer, P., & Al-Babili, S. (2012). The path from β-carotene to carlactone, a strigolactone-like plant hormone. *Science*, *335*(6074), 1348–1351. https://doi.org/10.1126/science.1218094

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings International Conference on Intelligent Systems for Molecular Biology*, *2*, 28–36.

Batra, R., Agarwal, P., Tyagi, S., Saini, D. K., Kumar, V., Kumar, A., Kumar, S., Balyan, H. S., Pandey, R., & Gupta, P. K. (2019). A study of CCD8 genes/proteins in seven monocots and eight dicots. *PLOS One*, *14*(3), 1–21. https://doi.org/10.1371/journal.pone.0213531

Buah, S. (2015). *Regulation of carotenoid biosynthesis in banana fruit* (Doctoral thesis, Queensland University of Technology, Australia). Queensland University of Technology ePrints. https://eprints.qut.edu.au/82719/

Calberto, G., Staver, G. C., & Siles, P. (2015). An assessment of global banana production and suitability under climate change scenarios. In A. Elbehri (Ed.), *Climate change and food systems: Global assessments and implications for food security and trade* (pp. 266-291). Food Agriculture Organization of the United Nations (FAO).

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2016). GenBank. *Nucleic Acids Research*, *44*(D1), D67-D72. https://doi.org/10.1093/nar/gkv1276

D'Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O., Noel, B., Bocs, S., Droc, G., Rouard, M., Da Silva, C., Jabbari, K., Cardi, C., Poulain, J., Souquet, M., Labadie, K., Jourda, C., Lengellé, J., Rodier-Goud, M., … Wincker, P. (2012). The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature*, *488*(7410), 213–217. https://doi.org/10.1038/nature11241

Davey, M. W., Gudimella, R., Harikrishna, J. A., Sin, L. W., Khalid, N., & Keulemans, J. (2013). "A draft *Musa balbisiana* genome sequence for molecular genetics in polyploid, inter- and intra-specific *Musa* hybrids". *BMC Genomics*, *14*(1), 683. https://doi.org/10.1186/1471-2164-14-683

Englberger, L., Wills, R. B. H., Blades, B., Dufficy, L., Daniells, J. W., & Coyne, T. (2006). Carotenoid content and flesh color of selected banana

cultivars growing in Australia. *Food and Nutrition Bulletin*, *27*(4), 281–291. https://doi.org/10.1177/156482650602700401

Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., & Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Research*, *8*(9), 967–974. https://doi.org/10.1101/gr.8.9.967

Hall, T. A. (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, *41*, 95–98.

Häuser, R., Pech, M., Kijek, J., Yamamoto, H., Titz, B., Naeve, F., Tovchigrechko, A., Yamamoto, K., Szaflarski, W., Takeuchi, N., Stellberger, T., Diefenbacher, M. E., Nierhaus, K. H., & Uetz, P. (2012). RsfA (YbeB) proteins are conserved ribosomal silencing factors. *PLOS Genetics*, *8*(7), e1002815. https://doi.org/10.1371/journal.pgen.1002815

Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics (Oxford, England)*, *17*(8), 754–755. https://doi.org/10.1093/bioinformatics/17.8.754

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A. R. N., Potter, S. C., Finn, R. D., & Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*, *47*(W1), W636–W641. https://doi.org/10.1093/nar/gkz268

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Wang, Z., Yamashita, R. A., Zhang, D., … Bryant, S. H. (2017). CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research*, *45*(D1), D200–D203. https://doi.org/10.1093/nar/gkw1129

Mathew, D., & Muhammed, T. A. (2015). Nutritional studies on lesser known wild edible bananas from Western Ghats, India. *Global Journal of Research on Medical Plants and Indigenous Medicine*, *3*(4), 127–133.

Mishra, S., Upadhyay, S., & Shukla, R. K. (2017). The role of strigolactones and their potential cross-talk under hostile ecological conditions in plants. *Frontiers in Physiology*, *7*, 691. https://doi.org/10.3389/fphys.2016.00691

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, *302*(1), 205–217. https://doi.org/10.1006/jmbi.2000.4042

Page, R. D. (1996). Tree View: An application to display phylogenetic trees on personal computers. *Bioinformatics*, *12*(4), 357–358. https://doi.org/10.1093/bioinformatics/12.4.357

Ronquist, F., & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, *19*(12), 1572–1574. https://doi.org/10.1093/bioinformatics/btg180

Solovyev, V., Kosarev, P., Seledsov, I., & Vorobyev, D. (2006). Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biology*, *7*(Suppl 1), S10. https://doi.org/10.1186/gb-2006-7-s1-s10

Wang, Y., Ding, G., Gu, T., Ding, J., & Li, Y. (2017). Bioinformatic and expression analyses

on carotenoid dioxygenase genes in fruit development and abiotic stress responses in *Fragaria vesca*. *Molecular Genetics and Genomics*, *292*(4), 895–907. https://doi.org/10.1007/s00438-017-1321-5

Waters, M. T., Brewer, P. B., Bussell, J. D., Smith, S. M., & Beveridge, C. A. (2012). The *Arabidopsis* ortholog of rice DWARF27 acts upstream of MAX1 in the control of plant development by strigolactones. *Plant Physiology*, *159*(3), 1073–1085. https://doi.org/10.1104/pp.112.196253