BIODIVERSITAS
Volume 22, Number 4, April 2021
Pages: 1829-1838

ISSN: 1412-033X
E-ISSN: 2085-4722
DOI: 10.13057/biodiv/d220428

Identification of conserved peptide upstream open reading frames (CPuORFs) in oil palm (*Elaeis guineensis*) genome

ANDREA PUTRI SUBROTO♥, REDI ADITAMA, ZULFIKAR ACHMAD TANJUNG, CONDRO UTOMO, TONY LIWANG

Department of Biotechnology, Plant Production and Biotechnology Division, PT SMART Tbk. Jl. Raya Cijayanti, Babakan Madang, Bogor, West Java, Indonesia. Tel. +62-21 392 5720, ♥email: biotechnology@sinarmas-agri.com

Manuscript received: 22 February 2021. Revision accepted: 19 March 2021.

Abstract. *Subroto AP, Aditama R, Tanjung ZA, Utomo C, Liwang T. 2021. Identification of conserved peptide upstream open reading frames (CPuORFs) in oil palm (*Elaeis guineensis*) genome. Biodiversitas 22: 1829-1838.* Upstream open reading frame (uORF) translation serves as one of several modulating mechanisms in gene expression. Conserved peptide uORF is a rare subset of uORF containing a peptide sequence that is conserved among eudicot families and negatively regulates the translation of its adjacent main ORF (mORF). This research aimed to identify oil palm CPuORFs by conducting a homology search between Arabidopsis (*Arabidopsis thaliana*) and rice (*Oryza sativa*) CPuORF databases against the existing oil palm (*Elaeis guineensis*) genome. Using alignment by Hidden Markov Model (HMM) approach, 48 oil palm CPuORFs were discovered which were later divided into 24 homology groups (HG) based on its gene ontology. The high degree of sequence homology indicates a high conservation CPuORFs among oil palm, Arabidopsis and rice. The CPuORFs peptide sequence conservation between oil palm progenies was later validated by sequencing of CPuORF22 and CPuORF29 as a reference, from several DxP Tenera crosses, individual Duras from different origins and American oil palm (*Elaeis oleifera*). This is the first study of oil palm CPuORF discovery which might provide a better understanding of gene expression regulation in oil palm.

Keywords: CPuORF, homology search, oil palm, translation regulation, uORF

Abbreviations: 5'UTR: Five prime untranslated region; uORF: upstream open reading frame; mORF: main ORF

INTRODUCTION

Gene expression needs to be regulated tightly in order to ensure a correct expression spatially, temporally, and also its dosage. There are several mechanisms of such regulation at different stages of expressions. A relatively new-discovery mechanism of gene regulation at translation level is the presence of conserved peptide upstream open reading frames (CPuORFs) (Hayden and Jorgensen 2007).

uORF is an open reading frame located at the 5'UTR of an mORF. A rare subset of uORF is called as conserved peptide upstream open reading frame (CPuORF) which negatively regulates the mORF by causing ribosome stalling and preventing ribosome re-initiation (Chang et al. 2000; Gaba et al. 2001; Rahmani et al. 2009; Van Der Horst et al. 2020). CPuORFs are typically comprised of 5-100 amino acids (aa) long and they are highly conserved at peptide level across eudicots (Hayden and Jorgensen 2007). The repression and peptide conservation of CPuORFs were confirmed through studies that introducing frameshift mutations to these conserved ORFs which alleviate their repression activity (Ebina et al. 2015; Noh et al. 2015; Takahashi et al. 2020). CPuORFs have been shown to regulate translation in response to abiotic signals and small molecules (Takahashi & Kakehi 2010; Alatorre-Cobos et al. 2012; Ivanov et al. 2018). Therefore, studying CPuORFs might help in understanding ribosome interaction with RNA to determine translation regulation.

The discovery of CPuORFs in plants was done commonly through motif search to the known full-length cDNA. Hayden and Jorgensen (2007) discovered CPuORF through comparison against rice (*Oryza sativa*) and Arabidopsis (*Arabidopsis thaliana*) full genome. These organisms were chosen because they diverged around 140-200 million years ago, and any trace of homologous uORFs at their mRNAs might signal the presence of conserved uORFs (Kumar et al. 2017). Twenty-six CPuORFs homology group based on gene ontology has been found by Hayden and Jorgensen (2007). From this point, there have been additional CPuORFs from other species being discovered (Tran et al. 2008; Jorgensen & Dorantes-Acosta 2012; Takahashi et al. 2012, 2020; Vaughn et al. 2012).

There have been no reports regarding oil palm (*Elaeis guineensis*) CPuORFs. As one of the important crop plants in the world, study about oil palm CPuORFs is important to set the basis of translation regulation in its genome. The advanced oil palm research, from the completion of oil palm genome (Singh et al. 2013; Ong et al. 2020) and several transcriptome data (Tranbarger et al. 2017) should provide enough data for discovering CPuORFs among its genes. This research aimed to discovery of oil palm CPuORFs through homology search against the previously (or publicly available) CPuORF using Hidden Markov

Model approach. According to our knowledge, this is the first study of CPuORFs in oil palm and we expect that it can open up to more insight towards gene expression in oil palm.

MATERIALS AND METHODS

CPuORFs screening in oil palm genome

The 5' UTR region of oil palm genome from National Center for Biotechnology Information (NCBI) database was mined. Whole-genome sequence and annotation file of *E. guineensis* were downloaded from NCBI GenBank with an accession number GCF_000442705.1. Position of 5'UTR of each gene was obtained by extracting the region between transcriptional start sites and start codon. The sequence of 5'UTR was extracted from genomics sequence based on 5'UTR position using BEDTools (Quinlan and Hall 2010).

CPuORF screening was conducted by aligning known CPuORFs (Hayden and Jorgensen 2007; Tran et al. 2008; Takahashi et al. 2012; Vaughn et al. 2012) to 5' UTR region of oil palm genome using HMMER software version 3. 1b2 (Finn et al. 2015). Conserved amino acid motifs were selected with E-value less than 10^{-5} as the threshold. The novel CPuORFs from oil palm were then aligned and re-clustered with the existing CPuORFs using Muscle Alignment software (Edgar 2004). CPuORFs homology group (HG) was based on the mORF gene ontology and the numbering referred to Hayden and Jorgensen (2007), Tran et al. (2008), Takahashi et al. (2012), and Vaughn et al. (2012). CPuORFs which were clustered in same HG, but located at the upstream of different mORFs were referred to as Elagu1, Elagu2, etc. The CPuORFs were given sequential ID numbers afterward.

Plant materials

Thirty-five samples were collected which comprised of 4 samples of Tenera Deli x LáMe (DL), 20 samples of Tenera Deli x AVROS (10 DA1 and 10 DA4 samples), 4 samples of Tenera Deli Dabou x LáMe (DDL), 4 samples of Tenera Yangambi selfing, and 3 samples of Dura Yangambi selfing. All samples were kindly provided by the Molecular Breeding section and Genomic and Transcriptomics section in Biotechnology Department, PT SMART Tbk.

Sequencing and alignment of CPuORFs between oil palm progenies

Amplification and sequencing of oil palm CPuORFs were conducted to verify the conservation of CPuORFs in several oil palm progenies. DNA extraction was performed using Isolate II Plant DNA kit (Bioline). CPuORF HG13 and HG17 Elagu1 (CPuORF22 and CPuORF 29, respectively) were chosen as references due to interesting function of their mORFs in oil palm. CPuORF22 and CPuORF29 were amplified from various oil palm leave samples through a PCR reaction. The PCR amplification was done using KOD plus FX neo (Toyobo) DNA polymerase kit, and a primer pair either of g13e1 F 5' CCATTTTCCGCAGTTGGACG 3' >< g13e1 R 5' CCACCAAAGCACACCGTCA 3' to amplify CPuORF22 or g17e1 5' CTCTTCTCCCAAGCTCCACC 3' >< g17e1 5' CGAGGGCTAGCGTTCTTCAT 3' to amplify CPuORF29. The primers were designed based on the CPuORF sequence using Primer3Plus (Untergasser et al. 2007). The amplicon was purified using Qiaquick PCR Purification Kit (Qiagen) and sequencing was conducted by the Firstbase Ltd., Singapore (IDT). All the 35 sample sequencing results were aligned and one additional sequence of *Elaeis oleifera* CPuORF from NCBI MPOB 08 scaffold was included to the alignment. The conservation of CPuORF22 and CPuORF 29 between 36 sequences were analyzed using Geneious software (Kearse et al. 2012).

RESULTS AND DISCUSSION

CPuORFs motif search in oil palm genome

Existing CPuORFs from public databases of previous studies were aligned and 48 CPuORFs were identified present in oil palm genome (Table 1). Initially, 90 aligned sequences were detected. Subsequently, sequences with E-value larger than 10^{-5} were eliminated, followed by elimination of sequences that do not have start codon (Table 2). This resulted in the discovery of 48 oil palm CPuORFs. These CPuORFs were found to be conserved within 15 HGs out of 26 HGs from Hayden and Jorgensen (2007), 5 out of 13 HGs from Takahashi et al. (2012) and 2 HGs out of 11 HGs and 4 HGs from Tran et al. (2008) and Vaughn et al. (2012), respectively. CPuORFs HG27-30 published by Jorgensen and Dorantes-Acosta (2012) were, unfortunately, unable to be screened in oil palm genome due to the alignment databases of those CPuORFs were unavailable. HG1 and HG9 have the most CPuORFs with 4 CPuORFs in each HG and there were 7 HGs that only have one CPuORF each (Table 1).

CPuORFs in oil palm ranging from 21 (CPuORF37) to 89 (CPuORF3) amino acids in length (Table 1). Some CPuORFs were conserved at the C-terminal, while varied at the N-terminal. This was shown by CPuORFs of HG1, HG4, HG11 and HG17. Contrary, some CPuORFs were conserved at the N-terminal and varied at the C-terminal, such as CPuORFs of HG6, HG7, HG9 and HG24n (Figure 1). The remaining CPuORF HGs were conserved throughout the CPuORF sequences (not shown).

We attempted to identify and predict the gene function of mORF attached to each CPuORF. CPuORFs exist at the upstream of transcription factor, translation factor, kinase, methyltransferase, and carboxylase indicating CPuORFs were highly represented in regulatory genes which confirms Hayden and Jorgensen (2007) as well as Jorgensen & Dorantes-Acosta (2012) findings (Table 2). Among those mORFs, HG13 and HG17 drove our attention. mORF of HG13 was involved in the methylation process of gene transcription, while mORF of HG17 was probably playing role during flowering process. Therefore, HG13 and HG17 in various oil palm progenies by deep sequencing analysis were further investigated.

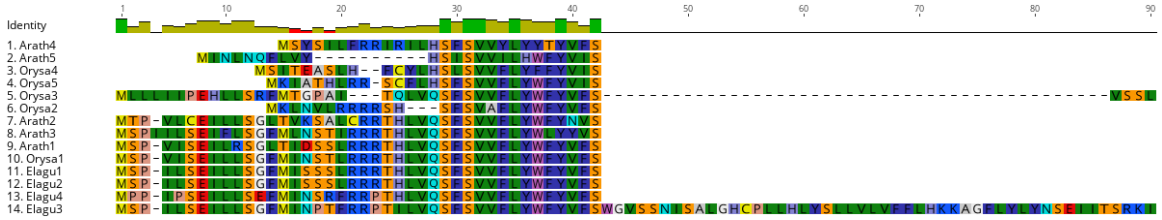
Table 1. CPuORF sequences identified in oil palm genome

Homology group	Alignment numbering[1]	CPuORFs sequence	CPuORF ID
Hayden and Jorgensen (2007)			
Group 1	Elagu1	MSPILSEILLSGFMISSSLRRRTHLVQSFSVVFLYWFYVFS	CPuORF1
	Elagu2	MSPILSEILLSGFMISSSLRRRTHLVQSFSVVFLYWFYVFS	CPuORF2
	Elagu3	MSPILSEILLSGFMINPTFRRPTILVQSFSVVFLYWFYVFSWGVSSNISALGHCPLLHL YSLLVLVFFLHKKAGFLYLYNSEIITSRKI	CPuORF3
	Elagu4	MPPIPSEILLSEFMINSRFRRPTHLVQSFSVVFLYWFYVFS	CPuORF4
Group 2	Elagu	MGCLLLGSTGPPIKRRAGLRRKQAGRGSYRGS	CPuORF5
Group 3	Elagu1	MESKGGKEKSSSSSSLQYEVPLGYSIEDVRPHGGIKKFQSAAYSNVSRRELRCSIFFFLFIC YLLFVY	CPuORF6
	Elagu2	MESKGGKKKSSSSSSLQYEVPLGYSIEDVRPHGGIKKFQSAAYSNVSRRELKCSIFCLSISF HWCTGY	CPuORF7
Group 4	Elagu	MGGGLGCQKTGRTATREALVRAYHISLLLSPVISFWDCIVRKIRYSFRPEWV	CPuORF8
Group 5	Elagu	MFLFSflvsSCSIvdSISSVLQNLRVFGPGLNPFAPFGMGNYSVSR	CPuORF9
Group 6	Elagu1	MSWFDKLPRSSNASLFMDGLMIFILHHSIPIRSSLDPIQS	CPuORF10
	Elagu2	MSWFDKLRRLSDASRYMDGLLIFILHHSIPIRLNLSLNPIQS	CPuORF11
Group 7	Elagu1	MSERDSFYLGLIGLLGEYCPTSRKKHIHSLDRVGCMHLEPFPDSALIQGCGWEQPFPLASDQ FCSEEASHSSSLTSASNPTSSIYYI	CPuORF12
	Elagu2	MSEQASLYLGSIGLLGGYSPTSRKRYRLPWERFGCMHQGQLQHSTSCCHMQPPLIAENLLGV LEGNGQTFSSGACYLS	CPuORF13
Group 9	Elagu1	MDRADTSGFVSGGSCCELRFDTYFHFIYISVD	CPuORF14
	Elagu2	MDRADTSGFVSGGGCNLTFHAYFHFTYSIVLSDKGLGMKH	CPuORF15
	Elagu3	MDRADTSGFVSGGgCCESRFVTYLDFIYLSAN	CPuORF16
Group 10.1[2]	Elagu1	MEQVPFWSSCFQCRVLLLQEVLDWRFFVLGDFLLISFVNCT	CPuORF17
	Elagu2	MEGAHSWSSCYQSKVSLFQEFFDWRFLPFGDFLIISFVNYTQWPAAFLCPTRAA	CPuORF18
	Elagu3	MERAHLWSSCYWHRVSLFQEALDQWFLAFGGFLLISFVNYT	CPuORF19
Group 11	Elagu1	MDCTHNCSDKKTLKRWFFIDKRVG	CPuORF20
	Elagu2	MDCTHDCSDKKTLKRWFFIDKRVG	CPuORF21
Group 13	Elagu1	MQQKGRSYNRRSRSFSRSRVAIEGA	CPuORF22
	Elagu2	MRQKGRFYNRSSRSFSRSRVAIEGS	CPuORF23
Group 14	Elagu1	MGFSLFPMKTSTRLLWSTSFFRHKIVVFF	CPuORF24
	Elagu2	MGFSLFPMKTSTRLLWSTSFFRHKIVVFF	CPuORF25
Group 15	Elagu1	MPWVPSFS-IryssSRKCVRLVVFFRVIV	CPuORF26
	Elagu2	MPWVSWTTNYfAECVRLAVFFRVIL	CPuORF27
	Elagu3	MLWLSLSVDVlRKVISLNLYCRVIL	CPuORF28
Group 17	Elagu1	MEAERMRQLMLLWIRSFRTRVALVGGNHTAARFCTR	CPuORF29
	Elagu2	MASVRVPARVWLPLQQTYRrLLVGGNHTSSRFCTR	CPuORF30
Group 19	Elagu1	MSNVPTSLCDSSLTLFQLAISSSDPWPFSF	CPuORF31
	Elagu2	MSNIPTSLCNSSLTLFRLAISRSDPWPFSL	CPuORF32
	Elagu 3	MSNIPTSLCNSSLTLFRLSISSSDPWPFSL	CPuORF33
	Elagu 4	MSHVPASLCDSSLTLSRLAISSSDPWPFSI	CPuORF34
Takahashi et al. (2012)			
AT2G27350	Elagu	MRIDRGMQRLCCGEESVLGWKKSFSRGIGIGAPKVRTRNKRMTKVWRL	CPuORF35
AT4G12790	Elagu	MDFSAFLSVYFAAIIWFGAYAVAIVISFRFF	CPuORF36
AT5G02480	Elagu	MGlISLIRLFPPVFGGFRGIL	CPuORF37
AT5G09330	Elagu1	MSFDIRKDYPWFLWCNFWFVCRSRLWCSRRILFRHWT	CPuORF38
	Elagu2	MSFDIRKDYLWFLWCDFWFVCRSRLWCSQRILFRHWT	CPuORF39
AT5G46590	Elagu	MILLHTSCILWKFCLVLE-KLSVEKSFYFWIFQIFNRYGVGFTMLILSKR	CPuORF40
Tran et al. (2008)			
AK072649	Elagu1	MNSRSTVAATVSCTSGNKGSFRILVCLLPGFLGAPVVDVNFRLFKAIVRIPSSSVVVSEICGGSQP	CPuORF41
	Elagu2	MNSRSTVAATVSCTSGNKGNFWILVCLLPVLLGAPVVYVNFHLFKAIVRIPSSSVVVSEICGGS	CPuORF42
AK072868_ uORF2[3]	Elagu1	MCLHAKESSSKVNDGVALSAMPLFILDEIEAQTYIQHVLHAESQHGSRNPQKAFVLDHDLGDDG	CPuORF43
	Elagu2	MCLHAKYFSSKVNDSVAPSELPLFILDEVGAQTYIQLMLHAQFLNGARNHQKAFVLDHDLVEDG	CPuORF44
	Elagu3	MCLNAKVSLSEANGIVAPFEPPLIHLYGIRQHTRIQFMLHVLMHNGDKRIIKPVVSDHDLGDDG	CPuORF45
Vaughn et al. (2012)			
Group24n	Elagu1	MLRRKPSKIEVKAEDREELEEVLRQRSlKAPRDGYHSANPNPNPLHKYLDPSPLDPAAKAQRIGLHQP	CPuORF46
	Elagu2	MLRPKPSKIEVKAEDREELEEALRQGSLKHPRGGQNPNPNPDPLHKYLDPNPTSKAQRIGLQKP	CPuORF47
Group25n	Elagu	MSQRPSVPHSSSIAFSLHSHLLVSSEMNPNSYWQQ	CPuORF48

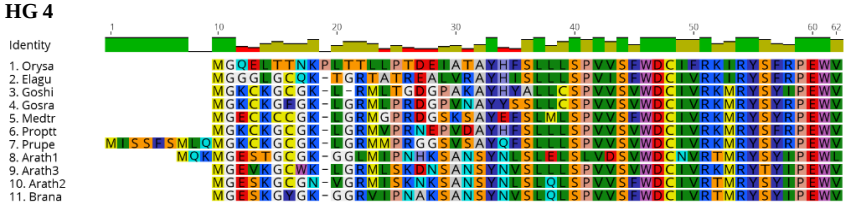
Note: [1]CPuORFs in same HG that located at different locus referred as Elagu1, Elagu2, etc. Single CPuORF in HG referred as Elagu. [2]Multiple uORF in one mORF referred as decimal as in Elagu 10.1 (Hayden and Jorgensen 2007). [3]Multiple uORF in one mORF referred as uORF1, uORF2, etc as in AK072868_uORF2 (Tran et al. 2008).

A. Conserved CPuORFs at the C-terminal and varied at the N-terminal

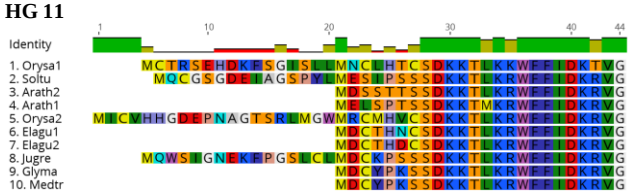
HG 1



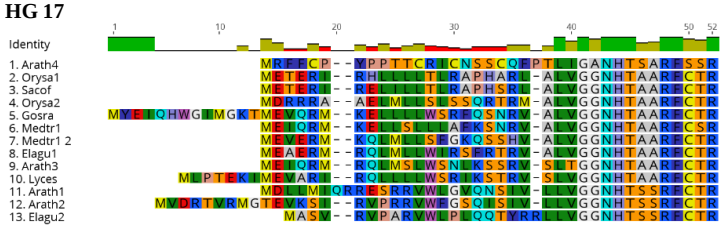
HG 4



HG 11

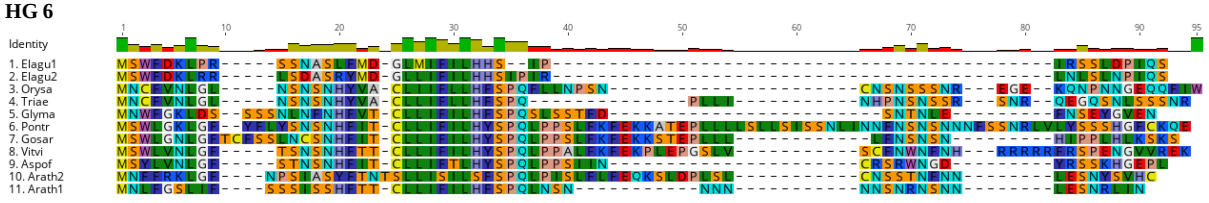


HG 17

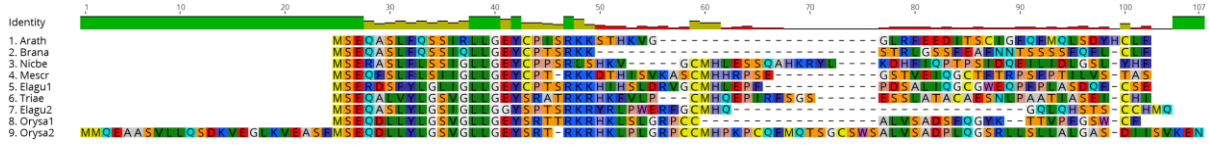


B. Conserved CPuORFs at the N-terminal and varied at the C-terminal

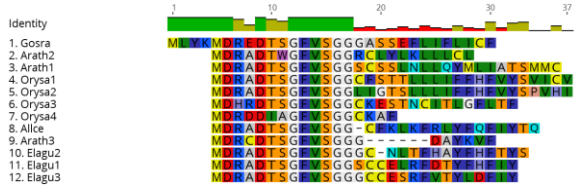
HG 6



HG 7



HG 9



HG24n

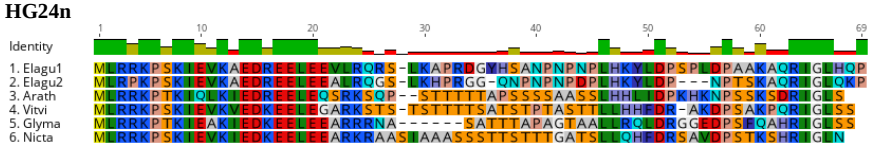


Figure 1. HMMER alignment of oil palm CPuORF (Elagu) from several HGs that showed conservation at different parts of the fragment. (A) CPuORFs were conserved at the C-terminal, while varied at the N-terminal. (B) CPuORFs were conserved at the N-terminal, while varied at the C-terminal

Validation of oil palm CPuORFs conservation

CPuORFs within HG13 and HG17 were analyzed by sequence alignment among different oil palm progenies. The nucleotide sequences were later translated into amino acid sequences. The sequencing result relatively showed no discrepancy between sample replications in each progeny. Variation within samples replication occurred within DA1 and DA4 samples during CPuORF22 sequencing, where the 3rd amino acid residue in DA1 sequence showed glutamine (Q) in eight samples and leucine (L) in two samples. While, within 10 DA1 and 10 DA4 samples, the 4th residue were varied where nine showed arginine (R) and one showed lysine (K), and four showed R and six were K, respectively (Figure 2). On the other hand, result of CPuORF29 sequencing showed no variation within sample replications (Figure 3). Therefore, only one sequence showed as a representative of each progeny (Figure 4). In case of variation within samples, the majority of amino acid residues were chosen as the representative. The amino acid sequence results between progenies showed that variation only occurred at the 4th residue of DA4 CPuORF22 samples (Figure 4A). At this point, R was replaced with K (Figure 4A, red box). The alignment showed highly conserved peptide sequence of CPuORFs between several oil palm progenies as well as *E. oleifera* (Elaol NCBI MPOB 08) (Figure 4).

The chemical features of CPuORFs HG13 and HG17 which could affect the secondary form of the nascent peptide were analyzed. On CPuORF22 within residue no.3 of DA1 samples, the amino acid variation has different chemical features, where L is non-polar and Q is polar. Meanwhile, on residue no.4, R and K have similar chemical features which are basic and contain amine group. Elagu2 of both HGs are mostly different from Elagu1. The Elagu2 of HG13 (CPuORF23) has 4 amino acids different at residue no.2, 7, 11, and 25. At residue no.2, Q was replaced with R. At residue no. 7, serine (S) which has polar, neutral, and hydrophilic was replaced with phenylalanine (F) which has aromatic and hydrophobic characteristics. At residue no. 11, R was replaced with S. Lastly, at residue 25 alanine (A) was replaced with S (Figure 4A). In HG17, Elagu2 (CPuORF30) showed large discrepancy when aligned to Elagu1 (CPuORF29). The methionine (M) position was aligned to position 5 since it had a shorter amino acid sequence than others. However, 15 residues of Elagu2 HG17 were identical to *A. thaliana* CPuORFs HG17. Interestingly, there was a different amino acid sequence at residue no 18, where R was replaced with S in *E. oleifera* sequence (Figure 4B).

Discussion

CPuORFs of oil palm was able to be identified with the availability of oil palm whole genome sequence (Singh et al. 2013). This study identified 48 CPuORFs in oil palm genome, which indicates the existence of conserved motifs in angiosperm uORFs. Oil palm and rice are monocots and included in Commelinids clade (Ma and Lu 2019). Taxonomically, oil palm and rice are categorized in different orders which are Arecales and Poales, respectively (Jouannic et al. 2005; Poczai and Hyvonen 2017). Both were separated around 114 mya according to the Timetree software (Kumar et al. 2017). *Arabidopsis* itself is a dicot. *Arabidopsis* and oil palm were separated around 160 mya according to the Timetree software (Kumar et al. 2017). The conservation of CPuORFs between oil palm, rice and *Arabidopsis* showed CPuORFs were conserved even between monocots and dicots (Hayden and Jorgensen 2007). Therefore, this peptide conservation is considered truly conserved, and not a result of gene duplication or sequence retention as reported by Takahashi et al. (2012).

Not all of the available HG databases were found in oil palm genome. Since HG20-26 were Arabidopsis paralogous conserved uORFS, they were less likely to be found in oil palm. Takahashi et al. (2012) screened CPuORFs by using broader EST databases from thousands of species which the identified CPuORFs might be taxonomically unique for several classes and therefore, could not be found in oil palm genome. Tran et al. (2008) published several HGs, some comprised of short conserved uORFs which were excluded in the alignment analysis. Those tiny uORFs were probably the artifact of point mutation that inserted in-frame start and/or stop codon in the 5'UTR.

BAIUCAS pipeline developed by Takahashi et al.(2012) established a threshold of CPuORFs length with minimum of 5 amino acids to be considered a functional peptide. Therefore, eight short conserved uORFs of Tran's database which were less than 5 residues, were not aligned to the oil palm genome. Three CPuORFs based on Tran's HGs were identical to the alignment result of Hayden and Jorgensen database. These are HG03-AK100589, HG10.1-AK069526 and HG11-AK103391. These redundant CPuORFs were included in Hayden and Jorgensen (2007) HG numbering in this study.

Other CPuORFs which were not found in oil palm might due to incomplete 5' UTR sequences. In 2013, oil palm genome was published with 1.8 gb genome sequence and predicted to contain 34,802 genes, thus allowing a comprehensive search of oil palm CPuORFs. However, around half of this genome dataset has not been unambiguously assembled, causing difficulties in discovering more CPuORFs. Oil palm, rice and Arabidopsis shared 40% homologous protein (Singh et al. 2013) but only 48 oil palm genes contain CPuORFs (less than 0.2%). This indicates that CPuORF is rare in oil palm genome and the gene expression regulation in oil palm depends less on this mechanism.

CPuORFs are over-represented in transcription factors and other gene regulators. Some CPuORFs have been studied extensively and the mORF showed to be translationally regulated by certain signals, like S1-group bZIP transcription factors (HG1), which respond to sucrose (Rahmani et al. 2009), S-adenosylmethionine decarboxylase (HG2) which response to polyamine (Ivanov et al. 2018), and phosphoethanolamine N-methyltransferase (HG13), which response phosphocholine (Alatorre-Cobos et al. 2012). Those signals showed to interact with CPuORFs nascent peptides and modulate mORFs translation. This translation control by CPuORFs might be interesting to be verified further in oil palm. CPuORFs which showed to response to

abiotic signals (Takahashi and Kakehi 2010; Fincato et al. 2011; Zhu et al. 2012) may allow a fast response to tackle abiotic stresses.

Two CPuORFs were chosen as reference to verify the conservation of CPuORFs between several *E. guineensis* progenies and also with *E. oleifera*. CPuORF HG13 was located at upstream of phosphoethanolamine N-methyltransferase 1, a key enzyme in phosphocholine (PCho) biosynthesis and probably plays an important role during methylation stage of a transcription process. CPuORFs of HG13 are S and R rich, comprising 50% of total peptide length (Hayden and Jorgensen 2007). uORFs with S rich motif were suggested to be able to change its RNA conformation through their phosphorylation which then promote/ inhibit ribosome stalling (Xiang et al. 2013) while R rich motifs can be involved in RNA binding (Bayer et al. 2005). CPuORF involvement in regulation of methyltransferase is also very interesting to be studied since one of oil palm challenge in increasing productivity is the formation of mantled fruit which was caused by reducing methylation in *MANTLED* gene (Ong-Abdullah et al. 2015). Therefore, understanding how CPuORFs control methylation might be worth to be studied.

The main ORF function of CPuORF HG 17 has not been reported. It was classified as alternative N-termini (aNT family aNT25) (Jorgensen and Dorantes-Acosta 2012). Protein sequence of CPuORF17's main ORF was blasted to NCBI database and the sequence was 50.38% homologs (query cover 93%) with flowering control protein like from *Actinidia chinensis* var. *chinensis*.

Table 2. Gene ontology of oil palm CPuORFs

HG	Alignment numbering	CPuORF ID	E-value	mORF	Locus	Gene ID
Group01	Elagu1	CPuORF1	7.3E-25	bZIP transcription factor 11-like	NC_026005	105056236
Group01	Elagu2	CPuORF2	7.3E-25	bZIP transcription factor 11-like	NC_026007	105058773
Group01	Elagu3	CPuORF3	8.9E-23	bZIP transcription factor 11-like	NC_025998	105047023
Group01	Elagu4	CPuORF4	6.4E-22	bZIP transcription factor 11-like	NC_025993	105046034
Group02	Elagu	CPuORF5	4.1E-13	transcription factor bHLH155 isoform X1	NC_025997	105044343
Group03	Elagu1	CPuORF6	2.2E-27	S-adenosylmethionine decarboxylase proenzyme-like	NC_026000	105050758
Group03	Elagu2	CPuORF7	3.1E-27	S-adenosylmethionine decarboxylase proenzyme	NC_025994	105039560
Group04	Elagu	CPuORF8	2.7E-26	uncharacterized protein LOC105059529	NC_026008	105059529
Group05	Elagu	CPuORF9	2.4E-16	ankyrin-1 isoform X1	NW_011552890	105035986
Group06	Elagu1	CPuORF10	0.00000032	probable polyamine oxidase 4 isoform X1	NC_026008	105058904
Group06	Elagu2	CPuORF11	0.000045	probable polyamine oxidase 4	NW_011551890	105035498
Group07	Elagu1	CPuORF12	7.7E-18	eukaryotic translation initiation factor 5-like	NW_011550984	105032882
Group07	Elagu2	CPuORF13	2.5E-17	eukaryotic translation initiation factor 5-like	NC_025993	105046943
Group09	Elagu1	CPuORF14	3.3E-14	uncharacterized protein LOC105058355	NC_026007	105058355
Group09	Elagu2	CPuORF15	5.4E-14	uncharacterized protein LOC105041692	NC_025995	105041692
Group09	Elagu3	CPuORF16	2.6E-13	uncharacterized protein LOC105056536	NC_026005	105056536
Group10.1	Elagu1	CPuORF17	4.2E-24	cyclin-dependent kinase F-4 isoform X1	NC_026008	105059272
Group10.1	Elagu2	CPuORF18	1.2E-18	cyclin-dependent kinase F-4-like isoform X2	NW_011551037	105033533
Group10.1	Elagu3	CPuORF19	6.7E-18	cyclin-dependent kinase F-4 isoform X2	NC_025996	105042519
Group11	Elagu1	CPuORF20	4.3E-18	probable trehalose-phosphate phosphatase F	NC_025996	105043601
Group11	Elagu2	CPuORF21	2E-17	probable trehalose-phosphate phosphatase F	NC_026003	105054029
Group13	Elagu1	CPuORF22	8.1E-10	phosphoethanolamine N-methyltransferase 1	NC_025993	105044829
Group13	Elagu2	CPuORF23	0.000000069	phosphoethanolamine N-methyltransferase 1-like	NC_025993	105044847
Group14	Elagu1	CPuORF24	4E-20	homeobox-leucine zipper protein HOX16	NC_025994	105039204
Group14	Elagu2	CPuORF25	4.4E-20	homeobox-leucine zipper protein HOX5-like	NC_026000	105050334
Group15	Elagu1	CPuORF26	2.2E-10	transcription factor bHLH144 isoform X2	NC_025998	105047231
Group15	Elagu2	CPuORF27	3.4E-09	transcription factor bHLH144-like	NC_025995	105040990
Group15	Elagu3	CPuORF28	0.00000014	transcription factor SAC51-like	NC_025995	105041410
Group17	Elagu1	CPuORF29	1.5E-21	uncharacterized protein LOC105058495	NC_026007	105058495
Group17	Elagu2	CPuORF30	1.7E-13	uncharacterized protein LOC109504882	NW_011550941	109504882
Group19	Elagu1	CPuORF31	2.8E-17	auxin-responsive protein SAUR32-like	NW_011551193	105034449
Group19	Elagu2	CPuORF32	3.9E-15	auxin-responsive protein SAUR32-like	NC_025996	105043524
Group19	Elagu 3	CPuORF33	2.4E-14	auxin-responsive protein SAUR32-like	NC_026003	105054051
Group19	Elagu 4	CPuORF34	4.1E-13	auxin-responsive protein SAUR32-like	NW_011551193	109505080
AT2G27350	Elagu	CPuORF35	7.80E-23	OTU domain-containing protein 5-B	NC_026001	105052100
AT4G12790	Elagu	CPuORF36	4.50E-07	GPN-loop GTPase 3	NC_026004	105055502
AT5G02480	Elagu	CPuORF37	9.60E-08	uncharacterized protein LOC105056958	NC_026006	105056958
AT5G09330	Elagu1	CPuORF38	8.00E-19	protein CUP-SHAPED COTYLEDON 3	NW_011552588	105037721
AT5G09330	Elagu2	CPuORF39	5.50E-19	NAC domain-containing protein 82-like	NC_025994	105035866
AT5G46590	Elagu	CPuORF40	3.60E-21	NAC domain-containing protein 72	NC_025996	105043851
AK072649	Elagu1	CPuORF41	3.80E-26	serine/threonine-protein kinase AtPK2/AtPK19	NC_025997	105044843
AK072649	Elagu2	CPuORF42	6.30E-24	serine/threonine-protein kinase AtPK2/AtPK19	NC_026006	105057217
AK072868 _uORF2	Elagu1	CPuORF43	1.70E-09	CBL-interacting protein kinase 18-like	NC_025997	105044796
AK072868 _uORF2	Elagu2	CPuORF44	2.50E-08	CBL-interacting protein kinase 18-like	NC_026006	105057189
AK072868 _uORF2	Elagu3	CPuORF45	2.90E-07	CBL-interacting protein kinase 18-like	NC_025997	105045953
Group24n	Elagu1	CPuORF46	1.40E-18	triphosphate tunel metalloenzyme 3-like	NC_025994	105039065
Group24n	Elagu2	CPuORF47	2.90E-14	triphosphate tunel metalloenzyme 3-like	NC_026000	105050110
Group25n	Elagu	CPuORF48	1.90E-21	mitochondrial uncoupling protein 5	NC_026007	105058492

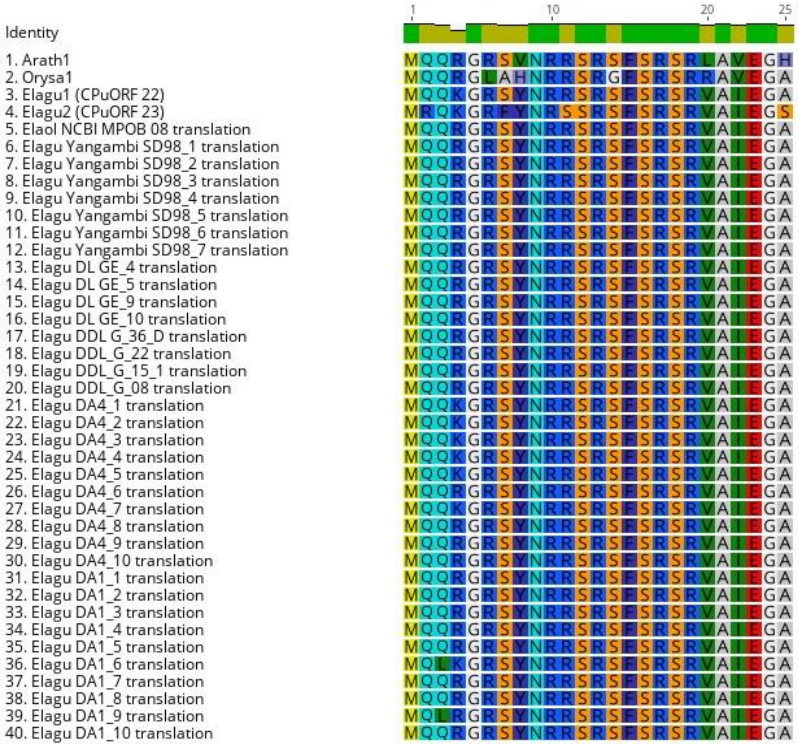


Figure 2. Amino acid sequence alignment of CPuORF HG13 Elagu1 (CPuORF22 in oil palm) from *A. thaliana* (Arath1), *O. sativa* (Orysa1), oil palm CPuORF22 and CPuORF23 from HMMER alignment, CPuORF22 from *E. oleifera* (Elaol NCBI MPOB 08 scaffold) and 35 sequencing results of oil palm CPuORF22.

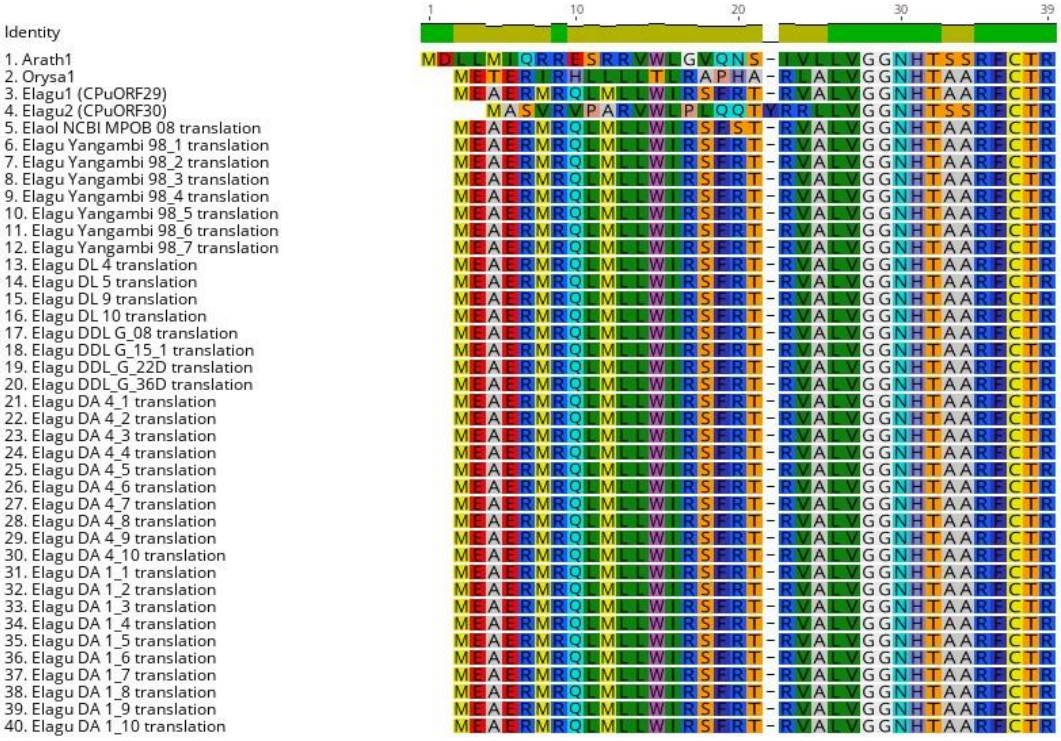


Figure 3. Amino acid sequence alignment of CPuORF HG17 Elagu1 (CPuORF29 in oil palm) from *A. thaliana* (Arath1), *O. sativa* (Orysa1), oil palm CPuORF29 and CPuORF30 from HMMER alignment, CPuORF29 from *E. oleifera* (Elaol NCBI MPOB 08 scaffold) and 35 sequencing results of oil palm CPuORF29.

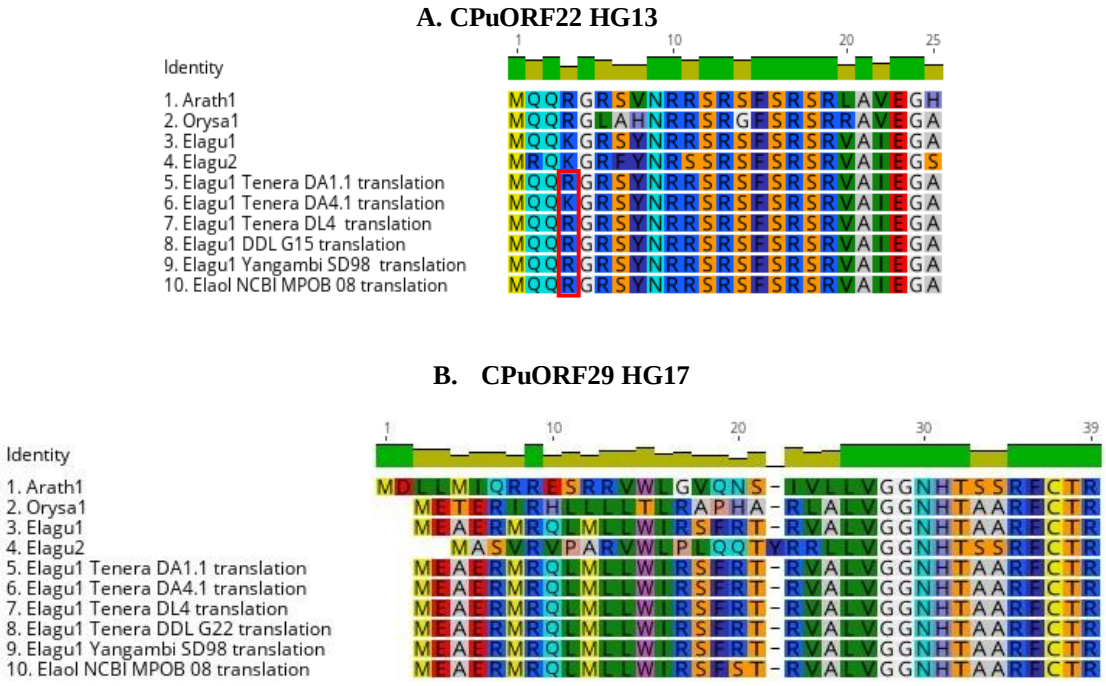


Figure 4. A. Alignment of CPuORFs22 from HG13, and B. CPuORFs29 from HG17 between several crossings of *E. guineensis* compared to *A. thaliana* (Arath1), *O. sativa* (Oryza1) and *E. oleifera* (Elaol NCBI MPOB 08 scaffold). Elagu1 and Elagu2 were CPuORFs found from alignment between previously published CPuORFs and oil palm genome. While Elagu1 DDL, Tenera DA1, Tenera DA4, Tenera DL and Yangambi were sequencing results of Elagu1 amplification from each CPuORF group. The red box highlighted variation in CPuORFs22 amino acid sequence between *E. guineensis* progenies

The alignment of both HGs showed that the CPuORFs are conserved in peptide level between several *E. guineensis* progenies, Arabidopsis, and rice (Figure 2, 3 and 4). Although most residues in most oil palm samples were homologous, some variations still occurred. This replacement of amino acid sequence might due to variation of its parental crossing (Paran and Zamir 2003). For now, we were not able to explain whether these substitutions will affect the CPuORFs regulation on its mORF translation, since CPuORFs between different species might contain regions that vary and conserved.

According to Vaughn et al. (2012), CPuORFs are divided into 2 groups, i.e. (1) CPuORFs that highly conserved at C-terminal with identical stop codon position and (2) CPuORFs in which the entire sequences or N-terminal and middle regions are conserved. Some suggest that ribosomes might stall at the conserved C-terminal of the CPuORFs (Takahashi et al. 2012), which did not explain how the translation regulation occurred in CPuORFs that conserved at the N-terminal only. The CPuORFs that conserved throughout the sequence might be translated into small peptides and cis-acting with ribosome to modulate translation (Araujo et al. 2012).

In conclusion, 48 oil palm CPuORFs have been successfully identified in oil palm genome based on existing CPuORFs databases. The regulation of each CPuORFs towards translation remains open questions. There are possibilities of more CPuORFs response to particular biochemistry signals. The founding of certain *A. thaliana* CPuORFs that respond to abiotic signals is a hint and interesting subject that needs to be revealed in oil palm. This study put fundamental database on CPuORF in oil palm for further advanced study.

ACKNOWLEDGEMENTS

This work was funded by PT SMART Tbk. under a research project code 3.3.4.068. We thank Shenni Maulina, Ahmad Jaelani, R. Gita Ismi Fauziah, and Raden Rizki Muhajir for their technical supports. We thank Reno Tryono, Roberdi and Victor Aprilyanto who read and improve the manuscript.

REFERENCES

Alatorre-Cobos F, Cruz-Ramirez A, Hayden CA,et al. 2012. Translational regulation of Arabidopsis XIPOTL1 is modulated by phosphocholine levels via the phylogenetically conserved upstream open reading frame 30. J Exp Bot 63 (14): 5203-5221. DOI: 10.1093/jxb/ers180.

Araujo PR, Yoon K, Ko D, et al. 2012. Before it gets started : Regulating translation at the 5 UTR. Comp Funct Genomics 2012: 475731. DOI: 10.1155/2012/475731.

Bayer TS, Booth LN, Knudsen SM, Ellington AD. 2005. Arginine-rich motifs present multiple interfaces for specific binding by RNA. RNA 11 12): 1848-1857. DOI: 10.1261/rna.2167605.

Chang KS, Lee SH, Hwang S Bin, Park KY. 2000. Characterization and translational regulation of the arginine decarboxylase gene in carnation (*Dianthus caryophyllus* L.). Plant J 24 (1): 45-56. DOI: 10.1046/j.0960-7412.2000.00854.x.

Ebina I, Takemoto-tsutsumi M, Watanabe S, et al. 2015. Identification of novel *Arabidopsis thaliana* upstream open reading frames that control

expression of the main coding sequences in a peptide sequence-dependent manner. Nucleic Acids Res 43 (3): 1562-1576. DOI: 10.1093/nar/gkv018.

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32 (5): 1792-1797. DOI: 10.1093/nar/gkh340.

Fincato P, Moschou PN, Spedaletti V, et al. 2011. Functional diversity inside the Arabidopsis polyamine oxidase gene family. J Exp Bot 62: 1155-1168.

Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, Bateman A, Eddy SR. 2015. HMMER web server: 2015 Update. Nucleic Acids Res 43: W30-W38. DOI: 10.1093/nar/gkv397.

Gaba A, Wang Z, Krishnamoorthy T, Hinnebusch AG, Sachs MS. 2001. Physical evidence for distinct mechanisms of translational control by upstream open reading frames. EMBO J 20: 6453-6463.

Hayden CA, Jorgensen RA. 2007. Identification of novel conserved peptide uORF homology groups in Arabidopsis and rice reveals ancient eukaryotic origin of select groups and preferential association with transcription factor-encoding genes. BMC Biol 5: 32. DOI: 10.1186/1741-7007-5-32.

Van Der Horst S, Snel B, Hanson J, Smeekens S. 2020. Novel pipeline identifies new upstream ORFs and non-AUG initiating main ORFs with conserved amino acid sequences in the 5' leader of mRNAs in. RNA 25 (3): 292-304. DOI: 10.1261/rna.067983.118 .

Ivanov IP, Shin B, Loughran G, et al. 2018. Polyamine control of translation elongation regulates start site selection on antizyme inhibitor mRNA via ribosome queuing. Mol Cell 70 (2): 254-264. DOI: 10.1016/j.molcel.2018.03.015.

Jorgensen RA, Dorantes-Acosta AE. 2012. Conserved peptide upstream open reading frames are associated with regulatory genes in Angiosperms. Front Plant Sci 3: 191. DOI: 10.3389/fpls.2012.00191.

Jouannic S, Argout X, Lechauve F, Fizames C, Borgel A, Morcillo F, Aberlenc-Bertossi F, Duval Y, Tregear J. 2005. Analysis of expressed sequence tags from oil palm (*Elaeis guineensis*). FEBS Lett 579 (12): 2709-2714. DOI: 10.1016/j.febslet.2005.03.093.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28 (12): 1647-1649. DOI: 10.1016/j.febslet.2005.03.093.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A Resource for timelines, time trees, and divergence times. Mol Biol Evol 34 (7): 1812-1819. DOI: 10.1093/molbev/msx116.

Ma Q, Lu Y. 2019. The complete chloroplast genome of Eichhornia crassipes (Pontederiaceae) and phylogeny of commelinids. Mitochondrial DNA B Resour 4: 3186-3187. DOI: 10.1080/23802359.2019.1667901.

Noh AL, Watanabe S, Takahashi H, Naito S, Onouchi H. 2015. An upstream open reading frame represses expression of a tomato homologue of Arabidopsis ANAC096, a NAC domain transcription factor gene, in a peptide sequence-dependent manner. Plant Biotechnol 32: 157-163. DOI: 10.5511/plantbiotechnology.15.0519a.

Ong-Abdullah M, Ordway JM, Jiang N, et al. 2015. Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. Nature 525: 533-537. DOI: 10.1038/nature15365.

Ong AL, Teh CK, Mayes S, Massawe F, Appleton DR, Kulaveerasingam H. 2020. An improved oil palm genome assembly as a valuable resource for crop improvement and comparative genomics in the Arecoideae subfamily. Plants 9 (11): 1476. DOI: 10.3390/plants9111476.

Paran I, Zamir D. 2003. Quantitative traits in plants: Beyond the QTL. Trends Genet 19: 303-306.

Poczai P, Hyvonen J. 2017. The complete chloroplast genome sequence of the CAM epiphyte Spanish moss (*Tillandsia usneoides*, *Bromeliaceae*) and its comparative analysis. PLoS One 12 (11): e0187199. DOI: 10.1371/journal.pone.0187199.

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26 (6): 841-842. DOI: 10.1093/bioinformatics/btq033.

Rahmani F, Hummel M, Schuurmans J, Wiese-Klinkenberg A, Smeekens S, Hanson J. 2009. Sucrose control of translation mediated by an upstream open reading frame-encoded peptide. Plant Physiol 150 (3): 1356-1367. DOI: 10.1104/pp.109.136036 .

Singh R, Low ETL, Ooi LCL, et al. 2013. The oil palm SHELL gene controls oil yield and encodes a homologue of SEEDSTICK. Nature 500: 340-344. DOI: 10.1038/nature12356.

Singh R, Ong-abdullah M, Low EL, et al. 2013. Oil palm genome sequence reveals divergence of interfertile species in Old and New Worlds. Nature 500 (7462): 335-339.DOI: 10.1038/nature12309.

Takahashi H, Hayashi N, Hiragori Y, Sasaki S, Motomura T, Yamashita Y, Naito S, Takahashi A, Fuse K, Satou K, Endo T, Kojima S,Onouchi H. 2020. Comprehensive genome-wide identification of angiosperm upstream ORFs with peptide sequences conserved in various taxonomic ranges using a novel pipeline, ESUCA. BMC Genomics 21 (1): 260. DOI: 10.1186/s12864-020-6662-5.

Takahashi H, Takahashi A, Naito S, Onouchi H. 2012. BAIUCAS: A novel BLAST-based algorithm for the identification of upstream open reading frames with conserved amino acid sequences and its application to the *Arabidopsis thaliana* genome. Bioinformatics 28: 2231–2241.

Takahashi T, Kakehi JI. 2010. Polyamines: Ubiquitous polycations with unique roles in growth and stress responses. Ann Bot 105 (1): 1-6. DOI: 10.1093/aob/mcp259.

Tran MK, Schultz CJ, Baumann U. 2008. Conserved upstream open reading frames in higher plants. BMC Genomics 9: 361. DOI: 10.1186/1471-2164-9-361.

Tranbarger TJ, Fooyontphanich K, Roongsattham P. 2017. Transcriptome analysis of cell wall and NAC domain transcription factor genes during *Elaeis guineensis* fruit ripening: Evidence for widespread conservation within monocot and eudicot lineages. Front Plant Sci 8: 603. DOI: 10.3389/fpls.2017.00603.

Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM. 2007. Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res 35: 71-74. DOI: 10.1093/nar/gkm306.

Vaughn JN, Ellingson SR, Mignone F, Von Arnim A. 2012. Known and novel post-transcriptional regulatory sequences are conserved across plant families. RNA 18 (3): 368-384. DOI: 10.1261/rna.031179.111.

Wang Z, Sachs MS. 1997. Ribosome stalling is responsible for arginine-specific translational attenuation in *Neurospora crassa*. Mol Cell Biol 17 (9): 4904-4913. DOI: 10.1128/mcb.17.9.4904.

Xiang S, Gapsys V, Kim HY, Bessonov S, Hsiao H, Möhlmann S, Klaukien V, Ficner R, Becker S, Urlaub H, Lührmann R, de Groot B, Zweckstetter M. 2013. Phosphorylation drives a dynamic switch in serine/arginine-rich proteins. Structure 21 (12): 2162-2174. DOI: 10.1016/j.str.2013.09.014.

Zhu X, Thalor SK, Takahashi Y, Berberich T, Kusano T. 2012. An inhibitory effect of the sequence-conserved upstream open-reading frame on the translation of the main open-reading frame of HsfB1 transcripts in Arabidopsis. Plant Cell Environ 35: 2014-2030.