BIODIVERSITAS
Volume 21, Number 11, November 2020
Pages: 5331-5337

ISSN: 1412-033X
E-ISSN: 2085-4722
DOI: 10.13057/biodiv/d211138

Analysis of codon usage bias reveals optimal codons in *Elaeis guineensis*

REDI ADITAMA[1], ZULFIKAR A. TANJUNG[1], WIDYARTINI M. SUDANIA[1], YOGO A. NUGROHO[2], CONDRO UTOMO[3,♥], TONY LIWANG[3]
[1]Section of Bioinformatics, SMART Biotechnology Center, PT SMART Tbk. Jl. Cijayanti, Babakan Madang, Bogor 16810, West Java, Indonesia
[2]Section of Molecular Breeding, SMART Biotechnology Center, PT SMART Tbk. Jl. Cijayanti, Babakan Madang, Bogor 16810, West Java, Indonesia
[3]Department of Biotechnology, Plant Production and Biotechnology Division, PT SMART Tbk. Jl. Cijayanti, Babakan Madang, Bogor 16810, West Java, Indonesia. Tel.: +62-21-3925720, ♥email: biotechnology@sinarmas-agri.com

Manuscript received: 14 August 2020. Revision accepted: 22 October 2020.

Abstract. *Aditama R, Tanjung ZA, Sudania WM, Nugroho YA, Utomo C, Liwang T. 2020. Analysis of codon usage bias reveals optimal codons in Elaeis guineensis. Biodiversitas 21: 5331-5337.* Codon usage bias of oil palm genome was reported employing several indices, including GC content, relative synonymous codon usage (RSCU), the effective number of codons (ENC), and codon adaptation index (CAI). Unimodal distribution of GC content was observed and matched with non-grass monocots characteristics. Correspondence analysis (COA) on synonymous codon usage bias showed that the main axis was strongly driven by GC content. The ENC and neutrality plot of oil palm genes indicating that natural selection played more vital role compared to mutational bias on shaping codon usage bias. A positive correlation between calculated CAI and experimental data of oil palm gene expression was detected indicating good ability of this index. Finally, eighteen codons were defined as "optimal codons" that may provide a useful reference for heterogeneous expression and genome editing studies.

Keywords: GC content, *Elaeis guineensis*, synonymous codons

INTRODUCTION

Codon usage bias (CUB) is a phenomenon where certain degenerative codons are used more frequently than expected by chance. It does not affect amino acids composition but has consequences on phenotype and fitness (Machado et al. 2017). The empirical pattern of codon usage is observed across species, within genomes and even on individual genes (Plotkin & Kudla 2011). The study of CUB on several plant species revealed the different patterns of codon usage between monocots and dicots, primarily on the usage of G/C on the third base position (Kawabe & Miyashita 2003).

The existence of CUB can be explained from three different points of views: the selection on codon usage (SCU) (Kliman 2014), mutational bias (MB) (Palidwor et al. 2010), and GC-biased gene conversion (gBGC) theory (Clément et al. 2017). The SCU theory suggests that CUB is maintained by selection due to its contribution to the efficiency and accuracy of amino acid sequences. The SCU theory has successfully described the codon usage pattern in *Drosophila pseudoobscura* (Kliman 2014), genus *Aspergillus* (Iriarte et al. 2012), and *Neurospora crassa* (Whittle et al. 2012). On the other hand, MB hypothesis suggests that CUB exists because of the non-randomness in mutation pattern, whereby some codons are mutated more frequently compared to others. The MB hypothesis has described codon usage patterns in Human and prokaryotes (Palidwor et al. 2010).

After years of debate over SCU and MB, gBGC hypothesis was evolved and gave a better explanation on codon usage and GC heterogeneity. The gBGC hypothesis was proposed based on the observation of the rapid increment of local GC content over genomic hotspots of recombination (Clément and Arndt 2013) and whole genome GC content over the high recombination rates of some species (Figuet et al. 2014; Weber et al. 2014). The gBGC hypothesis was successfully explained the codon usage patterns of some non-grass monocot species, including *Musa acuminata*, *Musa balbisiana*, *Phoenix dactylifera*, and *Spirodela polyrhiza* (Mazumdar et al. 2017).

African origin oil palm (*Elaeis guineensis* Jacq.) is the most important oil bearing-crop in the world which is capable to produce up to 12 t/ha/year vegetable oil (Corley and Tinker 2015). This crop has spread over tropical areas and has been cultivated in more than 16 countries (Wahid et al. 2015). Oil palm is one of the most important crops in Indonesia. It has been studied that oil palm cultivation has significant positive effects on farmer's livelihoods in Indonesia (Kubitza et al. 2018). In addition, national revenue from crude palm oil export has been increased in last few decades (Pacheco et al. 2017). However, the productivity of oil palm is still below the theoretically expected value, which is 18.5 t/ha/year (Woittiez et al. 2017). Many efforts have been made to increase oil palm productivity without expanding land use, including development of elite progenies through molecular breeding (Babu and Mathur 2016) and heterogeneous expression (Masani et al. 2018).

The information of CUB is essential in understanding molecular mechanism of gene expression as well as designing heterogeneous expression system. However, CUB in oil palm has not been studied extensively. The availability of whole genome sequence (Singh et al. 2013) and RNA-sequencing data (Lei et al. 2014; Ho et al. 2016; Othman et al. 2019) of oil palm enabled us to analyze CUB and optimal codons using bioinformatics tools. This study aims to analyze the codon composition of oil palm genome and present the pattern under several popular indices, including relative synonymous codon usage (RSCU), the effective number of codons (ENC), and codon adaptation index (CAI). These indices have been utilized extensively in studying codon usage in many organisms (Wang and Hickey 2007; Liu et al. 2010). Finally, the ultimate goal of this study is to find the "optimal codon" of oil palm that can be used in improving oil palm performance through molecular breeding and heterogeneous expression.

MATERIALS AND METHODS

Analysis pipeline

Figure 1 showed the bioinformatics pipeline to analyze CUB and optimal codons of oil palm. Whole genome sequence and annotation file of *E. guineensis* were downloaded from NCBI GenBank under accession number GCF_000442705.1. All of non-coding parts of the transcript including 5' and 3'UTR were removed from annotation file to make sure that only coding sequences were used in calculation. Coding sequences (CDS) were extracted from genome sequence employing BEDTools v2.26.0 (Quinlan and Hall 2010). For each CDS, sequences with more than one stop codon, lacking a start or stop codon and the length is not multiple of three were discarded. CDS were used for codon usage analysis using several indices, including GC content, GC content at first and second position of codon (GC12), GC content at third position of codon (GC3), relative synonymous codon usage (RSCU), and effective number of codon (ENC). Codon adaptation index (CAI) was analyzed and the result was used to determine optimal codons of oil palm.

GC content, GC12 and GC3

GC content, GC12 and GC3 of each CDS were calculated using General Codon Usage Analysis (GCUA) v1.2 (McInerney 1998). The distribution plot of GC content of oil palm genes was created using ggpubr R package (https://rpkgs.datanovia.com/ggpubr). Correlation between GC and GC3 was calculated using Pearson's coefficient and plotted to employ ggpubr. Neutrality plot was created by plotting GC3 against GC12 on a scatter plot using ggpubr. Linear regression analysis was calculated to determine the main influencing factors of codon bias.

Codon usage analysis

CUB patterns of oil palm were calculated using GCUA v1.2 (McInerney 1998) and presented on RSCU and ENC indices. RSCU value of each CDS was calculated using the formula described below (Sharp and Li 1987):

$$RSCU = \frac{g_{ij}}{\sum_{i}^{ni} g_{ij}} n_i$$

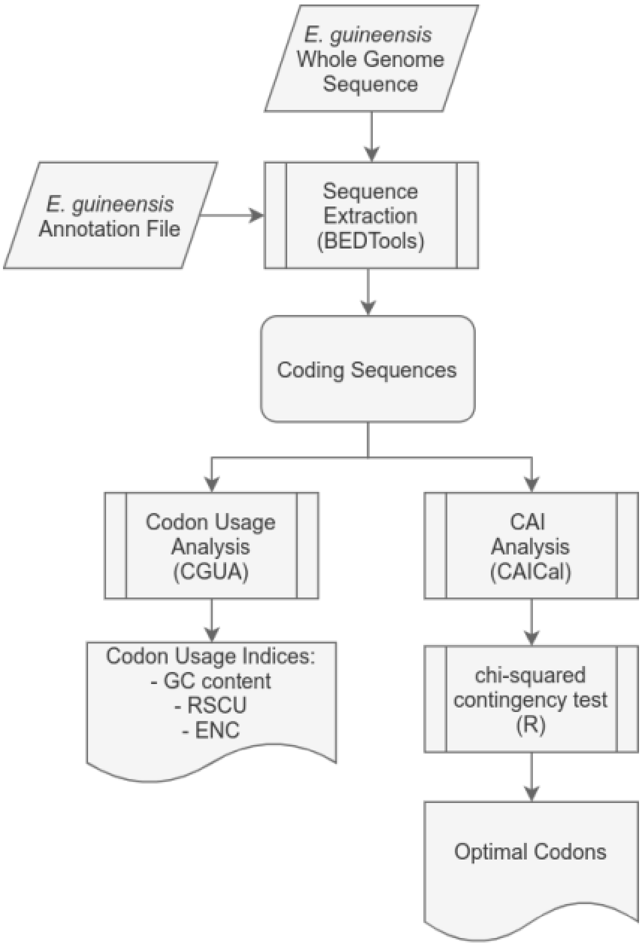
E. guineensis Whole Genome Sequence
E. guineensis Annotation File
Sequence Extraction (BEDTools)
Coding Sequences
Codon Usage Analysis (CGUA)
CAI Analysis (CAICal)
Codon Usage Indices:
- GC content
- RSCU
- ENC
chi-squared contingency test (R)
Optimal Codons

Figure 1. Bioinformatics pipeline for analyzing CUB and optimal codons in oil palm

Where: g_{ij} is the observed number of ith codon for the jth amino acid which has n_i kinds of synonymous codon. The RSCU value of 1 indicates no codon bias observed for those amino acids while RSCU more and less than 1 indicates positive and negative codon usage bias, respectively. To investigate the variation of RSCU values among genes, correspondence analysis (COA) was conducted using CGUA v1.2 and plotted using ggpubr.

To measure the distance between codon usages of a gene from equal usage, the effective number of codons (ENC) was calculated using the following formula (Ikemura 1981):

$$ENC = 2 + \frac{9}{\overline{F}_2} + \frac{1}{\overline{F}_3} + \frac{5}{\overline{F}_4} + \frac{3}{\overline{F}_6}$$

Where: $\overline{F}_k\ (k = 2,3,4,6)$ is the average homozygosity for the amino acids having degeneracy of k. ENC values range between 20 and 61. The lower value of ENC indicates more biased while higher value indicates less biased gene. The ENC plot was created by plotting GC3 against ENC for each gene using ggpubr.

Codon adaptation index

CAI quantifies the relative adaptiveness for each codon with respect to the codon usage of a reference set of highly expressed genes. This index was calculated using the following formula (Sharp and Li 1987):

$$CAI = exp\left(\frac{1}{L}\sum_{i=1}^{L} log\left(\omega_i(l)\right)\right), \omega_i = \frac{f_{ij}}{f_{xj}}$$

Where: ω_i is the relative adaptiveness of codon i, f_{ij} is the frequency of codon i encoding amino acid j, and L is the length of gene. CAI values range between 0 and 1 where the lower value suggests the random codon usage and tendency of lower expression while higher value suggests the extreme codon bias and the potential of a highly expressed gene. The CAI values for each oil palm gene were calculated using CAICal v1.4 (Puigbò et al. 2008) with 190 ribosomal protein-encoding genes of oil palm used as a reference for highly expressed genes.

Correlation with gene expression

To analyze the effect of codon adaptation index on gene expression, a set of RNA-seq data oil palm were downloaded from European Nucleotide Archive (ENA) under the accession number PRJEB7252. Raw reads of untreated root samples were cleaned and mapped to oil palm reference mRNA sequence using BWA (Li and Durbin 2009). Gene expression was quantified by eXpress (Roberts *et al.* 2011) using transcript per million (TPM) unit. The CAI value of each gene was plotted against gene expression using ggpubr. Pearson's correlation was used to calculate the correlation between CAI and gene expression.

Determination of optimal codon

The method to determine optimal codon was previously described and applied to *Zea mays* (Liu 2010). This method uses 5% of total genes with the highest and lowest CAI values as the high and low expression genes datasets. The 2×1 chi-squared contingency test was used to compare these two datasets where codons with frequency usage significantly higher ($P < 0.01$) in highly expressed genes than in low-level expression genes would be defined as optimal codons. The t-test was carried out to calculate the means of frequency usage in both high and low datasets.

RESULTS AND DISCUSSION

GC profile of genes

The distribution plot showed the unimodal distribution of oil palm genic GC content with the highest and lowest value is 27.82% and 73.74%, respectively (Figure 2A). The distribution is positively skewed with average value is 46.84% and the highest density GC content is about 42%. This observation is consistent with previous report on several monocot species (Clément al. 2014). Analysis employing Pearson's coefficient showed a relatively high correlation (R = 0.93) between GC content and GC3 of oil palm genes (Figure 2B).

Correspondence analysis

Correspondence analysis was conducted to examine the similarity of each gene in the oil palm in terms of RSCU (Figure 3). This analysis has reduced the complexity of 61 codons as variables for each gene into two primary axes. The distance between genes on the plot reflects their diversity of RSCU respected to two first axes. The result showed that the genes were more distributed on the first axis than the second. Most genes were distributed in -1.51 to 0.91 on the first axis and -0.91 to 0.91 on the second axis. To investigate the factor shaping the distribution, each gene on the plot was coded using different colors respected to GC content. Red, green, and blue colors indicated GC content below 45%, between 45% and 60%, and more than 60%, respectively. Each color category on the plot was separated along the first axis. The distribution of genes along second axis was different for each range of GC content. The genes with higher GC content were more distributed along second axis compared to genes with low GC content.

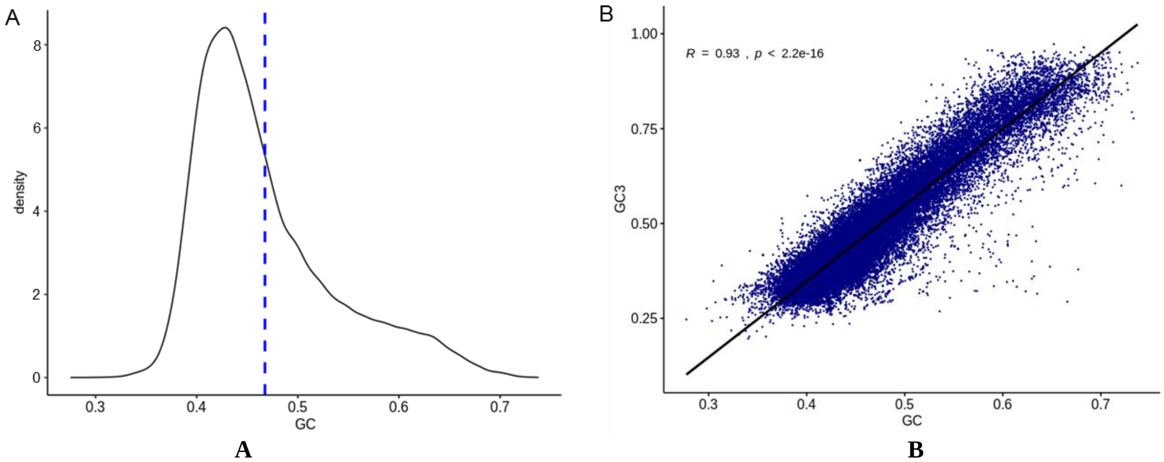

Figure 2. GC profile of oil palm genes: A. The distribution of GC content of oil palm gene (dashed blue line indicates the average number). B. Correlation between GC content and GC3

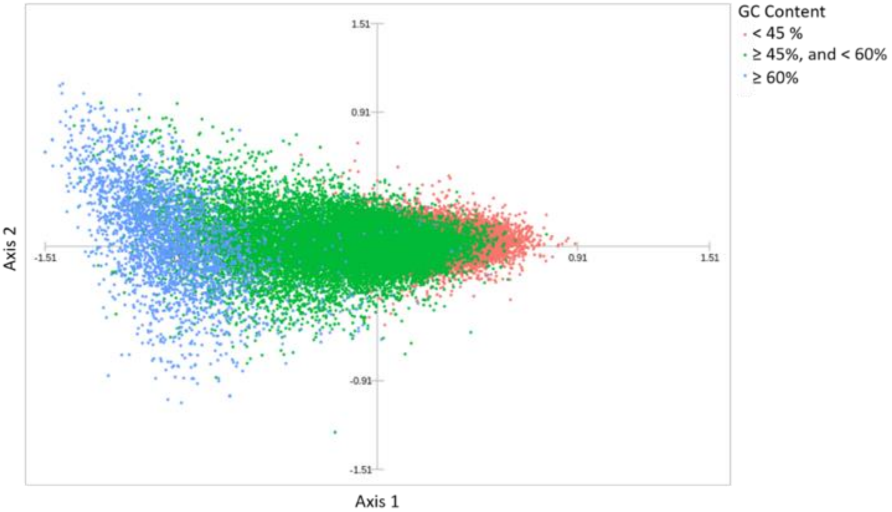

Figure 3. Correspondence analysis of codon usage of oil palm genes. The distribution of oil palm genes in terms of RSCU is defined by two axes. The red, green, and blue colors indicated GC content of less than 45%, equal or more than 45% but less than 60%, and equal or more than 60%, respectively

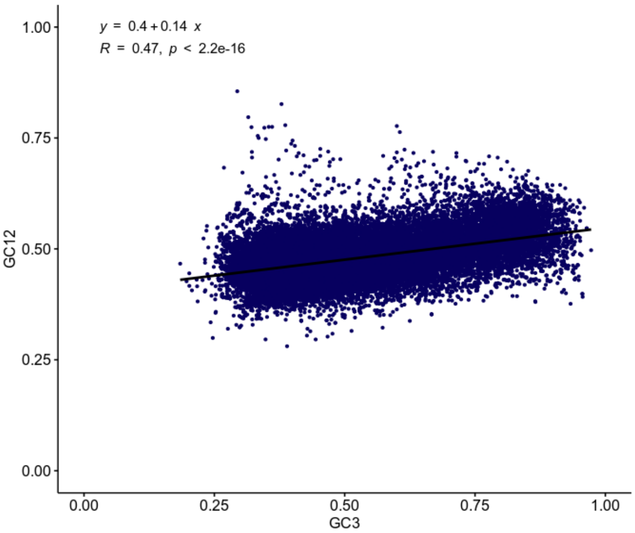

Figure 4. Neutrality plot analysis of oil palm genes

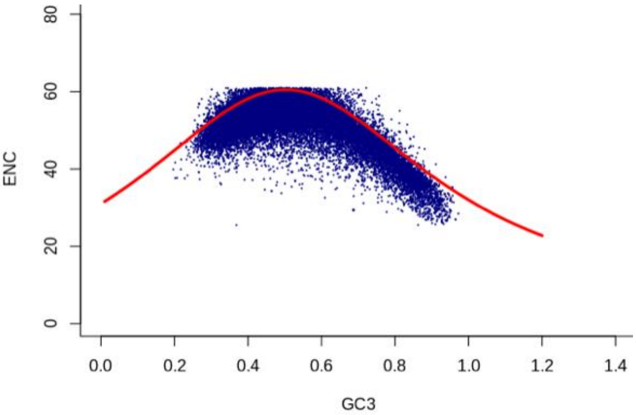

Figure 5. The relationship between GC3 and ENC for oil palm. The curve showed the expected value of ENC with respect to GC3.

Neutrality plot

Neutrality plot analysis of oil palm genes was conducted by plotting GC3 against GC12 of each gene. The result showed a significant and positive correlation between GC3 and GC12 (r = 0.47, p < 2.2e-16) (Figure 4). The range of GC3 was from 0.197 to 0.973 and the slope was 0.14. This indicated that the effect of mutational bias was only about 14%. This also indicated that natural selection played more vital role compared to mutational bias in shaping codon usage bias in oil palm.

The ENC plot

To examine the influence of GC3 on the codon usage, ENC for each gene were calculated and plotted against GC3 composition of individual gene (Figure 5). The continuous curve showed the expected value of ENC with respect to GC3 under the assumption that compositional constraints are the only determinant factor shaping the codon usage pattern. The plot showed that only small number of oil palm genes lay on expected curve. Most oil palm genes lay under expected value, suggesting that codon usage pattern is also affected by some other patterns, which are independent of compositional constraints. This also indicated that CUB of oil palm was mainly affected by natural selection rather than mutational bias. This result was consistent with neutrality plot analysis.

Correlation between CAI with gene expression

To study the correlation between codon usage and gene expression, CAI value was calculated for each gene. A set of 190 ribosomal protein encoded genes were used as a reference for highly expressed genes. CAI value for each gene was plotted against the experimental transcript abundance of oil palm seedling root under normal condition (Ho et al. 2016). A positive correlation (R = 0.32) was observed after expression value was transformed using logarithmic scale (Figure 6). The color code was used to distinguish the genes with low, medium, and high GC content. The red, green, and blue colors indicated GC content of less than 45%, equal or more than 45% but less than 60%, and equal or more than 60%, respectively. The result showed that the genes with high GC content were clearly separated with low GC content. The higher GC content, the higher CAI value, and wider range of value in experimental gene expression.

Translational optimal codons

The value of count and average RSCU of high as well as low expressed genes are listed (Table 1). Eighteen codons were determined as "optimal codons", which have significantly higher RSCU in high expressed compared to low expressed genes. All of the optimal codons are ended with G or C nucleotides.

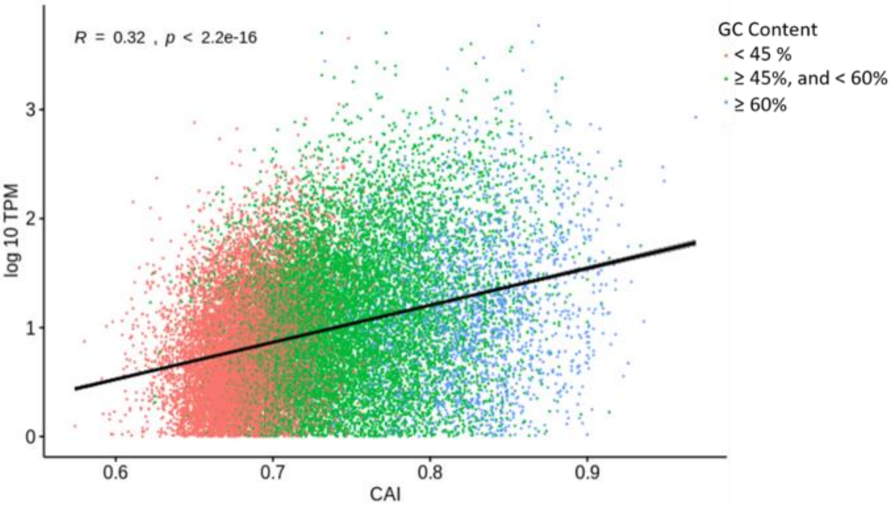

Figure 6. Correlation between CAI and experimental gene expression. The red, green, and blue colors indicated GC content of less than 45%, equal or more than 45% but less than 60%, and equal or more than 60%, respectively

Table 1. Translational optimal codons of oil palm

Amino acid	Codon	High		Low		Amino acid	Codon	High		Low	
		Count	RSCU	Count	RSCU			Count	RSCU	Count	RSCU
Phe	UUU	1,486	0.24	11,761	1.33	**Ser**	UCU	2,100	0.53	12,048	1.57
	UUC*	10,990	1.76	5,906	0.67		UCC*	12,730	3.21	4,927	0.64
Leu	UUA	246	0.05	9,502	1.05		UCA	761	0.19	12,948	1.68
	UUG	2,168	0.44	12,026	1.33		UCG	3,070	0.77	2,295	0.30
	CUU	2,817	0.57	12,817	1,41		AGU	703	0.18	8,059	1.05
	CUC*	18,706	3.79	4,757	0.53		AGC	4,433	1.12	5,833	0.76
	CUA	656	0.13	7,235	0.80	**Pro**	CCU	2,766	0.62	7,005	1.53
	CUG	4,987	1.01	8,015	0.88		CCC*	9,083	2.04	2,296	0.50
Ile	AUU	1,494	0.36	12,396	1.30		CCA	1,607	0.36	7,621	1.67
	AUC*	10,542	2.52	6,249	0.66		CCG*	4,321	0.97	1,355	0.30
	AUA	524	0.13	9,902	1.04	**Thr**	ACU	1,311	0.93	8,246	1.40
Val	GUU	1,773	0.36	10,655	1.49		ACC*	9,283	2.74	3,887	0.66
	GUC*	11,156	2.24	4,249	0.59		ACA	684	0.20	9,798	1.66
	GUA	469	0.09	6,572	0.92		ACG*	2,288	0.67	1,628	0.28
	GUG	6,530	1.31	7,100	0.99	**Ala**	GCU	3,389	0.47	12,292	1.54
Tyr	UAU	1,335	0.35	8,166	1.37		GCC*	18,927	2.61	4,393	0.55
	UAC*	6,343	1.65	3,727	0.63		GCA	1,523	0.21	13,399	1.68
His	CAU	1,324	0.37	9,609	1.51		GCG*	5,199	0.72	1,761	0.22
	CAC*	5,820	1.63	3,145	0.49	**Cys**	UGU	559	0.22	5,422	1.16
Gln	CAA	1,100	0.28	13,489	1.17		UGC	4,526	1.78	3,903	0.84
	CAG*	6,723	1.72	9,665	0.83	**Arg**	CGU	1,211	0.39	2,852	0.63
Asn	AAU	1,839	0.42	15,585	1.37		CGC*	9,060	2.88	1,497	0.33
	AAC	6,948	1.58	7,106	0.63		CGA	779	0.25	3,189	0.71
Lys	AAA	1,382	0.22	20,909	1.10		CGG*	3,944	1.25	2,273	0.51
	AAG	11,069	1.78	17,176	0.90		AGA	732	0.23	10,309	2.29
Asp	GAU	3,983	0.54	19,543	1.47		AGG	3,141	1.00	6,842	1.52
	GAC*	10,784	1.46	6,971	0.53	**Gly**	GGU	2,990	0.52	7,333	1.16
Glu	GAA	1,866	0.26	26,184	1.20		GGC*	12,396	2.15	3,941	0.62
	GAG	12,726	1.74	17,301	0.80		GGA	2,441	0.42	9,340	1.48
							GGG	5,216	0.91	4,683	0.74

Note: * optimal codon

Discussion

GC content is one of the most important factors in the evolution that shaping genomic structures and functions (Li et al. 2015). The distribution of oil palm genic GC content is unimodal that skewed to lower region (Figure 1.A). Unlike oil palm and other non-grass monocots, most of grass monocots have bimodal distribution of genic GC content (Tatarinova et al. 2010). Bimodality of genic GC content has been predicted to be a feature of ancestral monocots (Clément et al. 2014). The bimodality of ancestral monocots and most of grass is due to the strong gradient of GC content along the CDS in the 5' to 3' direction which makes the short intron-less genes rich in GC and long multi-exon genes have lower GC (Glémin et al. 2014). The unimodality on genic GC distribution of oil palm and other non-grass monocots is caused by the erosion of short GC-rich genes and the decrease in the 5' to 3' gradient. This mechanism can explain the skewness that occurred in oil palm genic GC content distribution.

A strong positive correlation between genic GC content and GC3 indicates that genic GC content is mainly driven by GC3 (Figure 1.B). This is because GC3 are less constrained by selection than first and second codon positions considering that most of variations on synonymous codon are located on the third position. The relationship between GC content and codon usage was also observed on correspondence analysis of RSCU where GC content drives the separation along the first axis (Figure 2). Since the genic GC content is driven by GC3 and in addition affects RSCU, it can be said that GC3 plays important role on the development of CUB. This is also observed in other species from both monocots and dicots, including wheat, barley, rice, maize, and *Arabidopsis* (Kawabe and Miyashita 2003)

To understand which factor that caused the development of CUB in oil palm, neutrality and ENC plot analysis have been conducted. Neutrality analysis showed that the effect of mutational bias was only 14% which indicated the dominance of selection-like force. On the other hand, ENC value of oil palm genes were under their expected value respective to GC3 which indicated that CUB was mainly affected by selection. The results of both neutrality and ENC analysis narrow down the factors to selection-like force, SCU or gBGC. Under SCU hypothesis, the bias of base and codon composition should be higher on highly expressed genes while according to gBGC hypothesis, the bias should be increased with recombination rates. Although the mechanisms of SCU and gBGC are clearly different, they are difficult to distinguish since they leave similar evolutionary footprints. Another study that analyzed the correlation analysis between several indices, including the direction of selection (DoS) against GC3 and expression level showed that gBGC is stronger than SCU (Clément et al. 2017).

Another approach to describe CUB of an organism is by using CAI. Unlike other indices, CAI employs a set of highly expressed genes as a reference to measure adaptiveness of codon. The higher the CAI of a gene, the more similar its CUB pattern compared to the references, which are highly expressed. Ribosomal protein encoded genes were used as reference for CAI calculation because they are the most stably expressed compared to other genes although some studies showed the tissue specificity of their expression (Thorrez et al. 2008). This may describe the positive but relatively weak correlation between CAI and transcript expression of root tissue of oil palm seedling (Figure 6). Since the genes are unlikely to be expressed across all tissues of an organism, the good reference will be dependent on the subset of tissues used in particular experiment. This is also observed in rice, where tissue-specific genes exhibit significantly different synonymous codon usage, although the effect is weak (Liu 2012).

The value of CAI was used on determining optimal codon in oil palm. The genes with highest CAI were considered as highly expressed while genes with lowest CAI were considered as lowly expressed. After statistical test was conducted on RSCU value of both highly and lowly expressed gene, eighteen G/C-ended codons were defined as optimal codons of oil palm. Although the tendency of highly expressed codon possess G/C at third position is similar, the number of optimal codon in oil palm is much lower compared to maize, which is 28 (Liu et al. 2010).

In heterologous protein expression, codon optimization has been widely used to enhance the expressivity (Zhao et al. 2010; Menzella 2011; Hu et al. 2013; Lanza et al. 2014). There are several online tools available for codon optimization, including GenSmart® Codon Optimization (https://www.genscript.com/tools/gensmart-codon-optimization) and IDT Codon Optimization Tool (https://www.idtdna.com/codonopt) but it would be better to use optimal codon rather than overall abundant codons (Liu et al. 2010). Optimal codons of oil palm later on can be used for heterogeneous expression system and with the advance of homologous recombination (HR) technology (Okamoto et al. 2019), can be used as a guide for modifying genes in order to enhance the protein expressivity.

Codon usage bias in oil palm was analyzed using several indices, including GC content, RSCU, ENC, and CAI. The unimodality of GC content distribution was observed and matched with non-grass monocots characteristics. Codon usage of oil palm is mainly driven by GC but also affected by some other factors which are independent of compositional constraint. Neutrality plot indicates that natural selection played more vital role compared to mutational bias on shaping codon usage bias. A positive correlation was observed between CAI, experimental transcript expression value, and GC content. Finally, eighteen codons were observed as "optimal codons" and can be used for codon optimization for many applications, such as heterogeneous expression and genome editing.

ACKNOWLEDGEMENTS

We thank the management of PT SMART Tbk for supporting this research and publication. Credits also addressed to Reno Tryono, Marcelinus R. Hatorangan and Victor Aprilyanto for their suggestion for the manuscript.

REFERENCES

Babu BK, Mathur RK. 2016. Molecular breeding in oil palm (*Elaeis guineensis*): Status and future perspectives. Prog Holt 48 (2): 123-131.

Clément Y, Arndt PF. 2013. Meiotic recombination strongly influences GC-content evolution in short regions in the mouse genome. Mol Biol Evol 30 (12): 2612-2618.

Clément Y, Fustier MA, Nabholz B, Glemin S. 2014. The bimodal distribution of genic GC content is ancestral to monocot species. Genome Biol Evol 7 (1): 336-348. DOI: 10.1093/gbe/evu278

Clément Y, Sarah G, Yan Holtz, Homa F, et al. 2017. Evolutionary forces affecting synonymous variations in plant genomes. PLoS Genet 13 (5): e1006799. DOI: 10.1371/journal.pgen.1006799

Corley RHV, Tinker PB. 2015. The Oil Palm. John Wiley & Sons, Chicester.

Figuet E, Ballenghien M, Romiguier J, Galtier N. 2014. Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. Genome Biol Evol 7 (1): 240-250. DOI: 10.1093/gbe/evu277

Glémin S, Clément Y, David J, Ressayre A. 2014. GC content evolution in coding regions of angiosperm genomes: A unifying hypothesis. Trends Genet 30 (7): 263-270. DOI: 10.1016/j.tig.2014.05.002

Ho CL, Tan YC, Yeoh KA, Ghazali AK, Yee WY, Hoh CC. 2016. *De novo* transcriptome analyses of host-fungal interactions in oil palm (*Elaeis guineensis* Jacq). BMC Genomics 17: 66. DOI: 10.1186/s12864-016-2368-0

Hu H, Gao J, He J, Yu B, Zheng P, Huang Z, Mao X, Yu J, Han G, Chen D. 2013. Codon optimization significantly improves the expression level of a keratinase gene in *Pichia pastoris*. PLoS One 8 (3): e58393. DOI: 10.1371/journal.pone.0058393

Iriarte A, Sanguinetti M, Fernandez-Calero T, Naya H, Ramon A, Musto H. 2012. Translational selection on codon usage in the genus *Aspergillus*. Gene 506 (1): 98-105. DOI: 10.1016/j.gene.2012.06.027

Ikemura T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translation system. J Mol Biol 151 (3): 389-409. DOI: 10.1016/0022-2836(81)90003-6

Kawabe A, Miyashita NT. 2003. Patterns of codon usage bias in three dicot and four monocot plant species. Genes Genet Syst 78: 343-352.

Kliman RM. 2014. Evidence that natural selection on codon usage in *Drosophila pseudoobscura* varies across codons. G3 (Bethesda) 4 (4): 681-692. DOI: 10.1534/g3.114.010488

Kubitza C, Krishna VV, Alamsyah Z, Qaim M. 2018. The economics behind and ecological crisis: Livelihood effects of oil palm expansion in Sumatra, Indonesia. Hum Ecol 46: 107-116. DOI: 10.1007/s10745-017-9965-7

Lanza AM, Curran KA, Rey LG, Alper HS. 2014. A condition-specific codon optimization approach for improved heterologous gene-expression in *Saccharomyces cerevisiae*. BMC Syst Biol 8: 33. DOI: 10.1186/1752-0509-8-33

Lei X, Xiao Y, Xia W, Mason AS, Yang Y, Ma Z, Peng M. 2014. RNA-seq analysis of oil palm under cold stress reveals a different C-repeat binding factor (CBF) mediated gene expression pattern in *Elaeis guineensis* compared to other species. PLoS One 9 (12): e114482. DOI: 10.1371/journal.pone.0114482

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25 (14): 1754-1760.

Li J, Zhou J, Wu Y, Yang S, Tian D. 2015. GC-content of synonymous codons profoundly influences amino acid usage. G3 5 (10): 2027-2036. DOI: 10.1534/g3.115.019877

Liu H, He R, Zhang H, Huang Y, Tian M, Zhang J. 2010. Analysis of synonymous codon usage in *Zea mays*. Mol Biol Rep 37 (2): 677-684. DOI: 10.1007/s11033-009-9521-7

Liu Q. 2012. Mutational bias and translational selection shaping the codon usage pattern of tissue-specific genes in rice. PLoS One 7 (10): e48295. DOI: 10.1371/journal.pone.0048295

Machado HE, Lawrie DS, Petrov DA. 2017. Strong purifying selection on codon usage bias. bioRxiv. DOI: 10.1101/106476

Masani MYA, Izawati AMD, Rasid, OA, Parveez GKA. 2018. Biotechnology of oil palm: Current status of oil palm genetic transformation. Biocatal Agric Biotechol 15: 335-347. DOI: 10.1016/j.bcab.2018.07.008

Mazumdar P, Othman RYB, Mebus K, Ramakrishnan N, Harikrishna JA. 2017. Codon usage and codon pair patterns in non-grass monocot genomes. Ann Bot 120 (6): 893-909. DOI: 10.1093/aob/mcx112

McInerney JO. 1998. GCUA: General codon usage analysis. Bioinformatics 14 (4): 372-373. DOI: 10.1093/bioinformatics/14.4.372

Menzella HG. 2011. Comparison of two codon optimization strategies to enhance recombinant protein production in *Escherichia coli*. Microb Cell Fact 10: 15. DOI: 10.1186/1475-2859-10-15

Okamoto S, Amaishi Y, Maki I, Enoki T, Mineno J. 2019. Highly efficient genome editing for single-base substituting using optimized ssODNs with Cas9-RNPs. Sci Rep 9: 4811. DOI: 10.1038/s41598-019-41121-4

Othman NQ, Sulaiman S, Lee YP, Tan JS. 2019. Transcriptomic data of mature oil palm basal trunk tissue infected with *Ganoderma boninense*. Data Brief 25: 104288. DOI: 10.1016/j.dib.2019.104288

Pacheco P, Gnych S, Dermawan A, Komarudin H, Okarda B. 2017. The palm oil global value chain: Implications for economic growth and social and environmental sustainability. Working Paper 220. Bogor, Indonesia: CIFOR.

Palidwor GA, Perkins TJ, Xia X. 2010. A general model of codon bias due to GC mutational bias. PLoS One 5 (10): e13431. DOI: 10.1371/journal.pone.0013431

Plotkin J, Kudla G. 2011. Synonymous but not the same: The causes and consequences of codon bias. Nat Rev Genet 12 (1): 32-42. DOI: 10.1038/nrg2899

Puigbò P, Bravo IG, Garcia-Vallve S. 2008. CAIcal: A combined set of tools to assess codon usage adaptaion. Biol Direct 3: 38. DOI: 10.1186/1745-6150-3-38

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

Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. 2011. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol 12 (3): R22. DOI: 10.1186/gb-2011-12-3-r22

Sharp PM, Li WH. 1987. The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential application. Nucleic Acids Res 15 (3): 1281-1295. DOI: 10.1093/nar/15.3.1281

Singh R, Ong-Abdullah M, Low EL, Manaf MAA, Rosli R, 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

Tatarinova TV, Alexandrov NN, Bouck JB, Feldmann KA. 2010. GC3 biology in corn, rice, sorghum and other grasses. BMC Genomics 11 (1): 308. DOI: 10.1186/1471-2164-11-308

Thorrez L, Van Deun K, Tranchevent LC, Lommel LV, Engelen K, Marchal K, Moreau Y, Mechelen I, Schuit F. 2008. Using ribosomal protein as reference: A tale of caution. PLoS One 3 (3): e1854. DOI: 10.1371/journal.pone.0001854

Wahid MB, Abdullah SNA, Henson IE. 2005. Oil Palm – Achievements and potential. Plant Prod Sci 8: 288-297. DOI: 10.1626/pps.8.288

Wang CH, Hickey DA. 2007. Rapid divergence of codon usage patterns within the rice genome. BMC Evol Biol 7: S6. DOI: 10.1186/1471-2148-7-S1-S6

Weber CC, Boussau B, Romiguier J, Jarvis ED, Ellegren HE. 2014. Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. Genome Biol 15 (12): 549. DOI: 10.1186/s13059-014-0549-1

Whittle CA, Sun Y, Hohanneson H. 2012. Genome-wide selection on codon usage at the population level in the fungal model organism *Neurospora crassa*. Mol Biol Evol 29 (8): 1975-1986. DOI: 10.1093/molbev/mss065

Woittiez LS, van Wijk MT, Slingerland M, van Noordwijk M. 2017. Yield gaps in oil palm: A quantitative review of contributing factors. Europ J Agronomy 83: 57-77. DOI: 10.1016/j.eja.2016.11.002

Zhao S, Huang J, Zhang C, Deng L, Hu N, Liang Y. 2010. High-level expression of an *Aspergillus niger* Endo-B-1,4-Glucanase in *Pichia pastoris* through gene codon optimization and synthesis. J Microbiol Biotechnol 20 (3): 467-473.