IOP Conference Series: Earth and Environmental Science

PAPER • OPEN ACCESS

Genome-wide SNP-discovery and analysis of genetic diversity in oil palm using double digest restriction site associated DNA sequencing

To cite this article: Y A Nugroho *et al* 2019 *IOP Conf. Ser.: Earth Environ. Sci.* **293** 012041

View the article online for updates and enhancements.

This content was downloaded from IP address 120.78.68.241 on 11/10/2019 at 11:10

Genome-wide SNP-discovery and analysis of genetic diversity in oil palm using double digest restriction site associated DNA sequencing

Y A Nugroho[1,*], Z A Tanjung[2], D Yono[1], A S Mulyana[1], H M Simbolon[3], A S Ardi[4], Y Y Yong[4], C Utomo[5] and T Liwang[5]

[1]Section of Molecular Breeding, Plant Production and Biotechnology Division, PT SMART, Tbk, Bogor, 16810, Indonesia.
[2]Section of Bioinformatics, Plant Production and Biotechnology Division, PT SMART, Tbk, Bogor, 16810, Indonesia
[3]Dami Mas Sejahtera, Plant Production and Biotechnology Division, PT SMART, Tbk, Riau, 28464, Indonesia
[4]SMART Research Institute, Jalan Teuku Umar 19, Pekanbaru, Riau, 28141, Indonesia
[5]Plant Production and Biotechnology Division, PT SMART Tbk, Bogor, 16810, Indonesia

*Corresponding author: biotechnology@sinarmas-agri.com

Abstract. The oil palm conventional breeding is faced with many challenges, including the long breeding cycle and limited genetic variations available from the existing breeding populations. Recent genomic research revealed the genetic properties of various oil palm populations and could be exploited by recombining favorable alleles through molecular breeding and selection. In aimed to discover high quality single nucleotide polymorphisms (SNPs) sampled from 236 diverse genetic backgrounds palms, a paired-end ddRAD-sequencing was employed to discover allelic variations. Whole 195.62 Gb clean data were generated using ddRAD-seq. After stringent filtering, 8 189 SNPs in the annotated genes were discovered based on the oil palm EG05 reference genome. The high polymorphism rate was observed across the assayed populations, whereby 29 % of loci showed moderate to high putative impact on amino acid changes in silico as predicted by SnpEff software. The Angola and Deli origins seemed to have undergone inbreeding as reflected by the fixation index and the lower observed heterozygosity compared to expected heterozygosity. Assessment of genetic stratification and structure was not strong enough to differentiate the core breeding materials according to their prior genetic backgrounds. The results demonstrated the robustness and high-throuputness of ddRAD sequencing method for SNP discovery.

Keywords: *Eco*RI, *Elaeis guineensis, Mse*I, oil palm, SnpEff

1. Introduction

High oil quantity and quality are the main goals in oil palm cultivation and breeding programs. A breeding program to improve oil palm yield has a long history, dating back to the late 19th century. The Deli dura base population, descendants of four palms first introduced in the Bogor Botanical Gardens (Indonesia) in 1848 and AVROS breeding populations were the base populations used by most oil palm breeders worldwide [1]. These breeding populations of restricted origins have

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

undergone more than five generations of breeding improvements. As such, both the Deli and AVROS breeding populations were highly inbred [2]. Further progress through conventional breeding and selection for yield improvement may difficult or not be significant due to the limited allelic variability available within these core populations. Introgression of other germplasms into the current breeding populations will be necessary to introduce new favorable and desirable gene alleles.

The intrapopulation sib-matting, and specific trait selection performed by oil palm research centers led to the development of several specific sub-populations with unique genetic properties although they were descended from the same ancestor's several generations back. These advanced breeding populations are very inbred and genetic characterization is required to evaluate and identify the gene pools for the conservation of core breeding materials or utilize it for introgression of new favorable allelic variant from other populations.

Population molecular diversity studies have become standard practice in many oil palm breeding program [3]. They are important for developing strategies for germplasm prospection, organization, and development of core collections. Such information will facilitate the selection, maintenance, and exploitation of key genotypes and populations for further development of heterotic populations. The recent genomic study has been established for the conservation of oil palm core collections consisting of a limited number of accessions with the maximum genetic diversity and representing the widest genetic backgrounds contained within the entire collections [4, 5].

Double digest Restriction Associated DNA (ddRAD) sequencing is one of the genotyping techniques which rely on restriction enzymes to determine the set of loci to be sequenced. It provides a high-throughput technique that is commonly used to uncover allelic variation across the genome in a single, simple and cost-effective method [6, 7] in many species, including the oil palm [8]. The genetic information generated from this technique can directly be used by breeders to scan the available breeding stocks for genetic diversity study for investigating the population structure and differences among collections. Moreover, genome-wide association study to identify the genic region controlling particular traits of interest will allow the breeders to collect the prolific loci into elite progenies through molecular marker-assisted breeding. In this study, the ddRAD-seq technology was employed to perform an oil palm genome-wide SNP discovery for core breeding materials. A collection set of SNP probes have been developed for mass genotyping in the near future.

2. Methodology

2.1. Planting material and genomic DNA extraction

A total of 236 palms comprising of five origins of *Elaeis guineensis*, nine introgression crosses, and an interspecific cross between *E. oleifera* and *E. guineensis* (table 1) were used for genotyping analysis. The *E. guineensis* samples consisted of 57 palms of undomesticated germplasm (Angola and Cameroon origin), 84 inbred lines of advanced breeding materials (Deli, AVROS, and Yangambi) and 88 introgressed palms from eight diverse genetic backgrounds.

Total genomic DNA was isolated from liquid Nitrogen ground leaf tissues using the NucleoSpin® Plant II mini kit (Macherey-Nagel, Germany) in accordance with the manufacturer’s protocols. DNA concentration and purity were estimated by the NanoDrop® 2000 UV-Vis Spectrophotometers (Thermo Scientific, USA). The genomic DNA samples were checked for total volume, concentration, and DNA integrity according to the Beijing Genome Institute (BGI) specifications. Samples meeting the requirement were then processed into library preparation for double-digested-RAD sequencing. The ddRAD sequencing was performed by the Beijing Genome Institute in Hong Kong, China.

The genomic DNA was digested with both *Eco*RI (recognition site 5’-G/AATTC-3’) and *Mse*I (recognition site 5’-T/TAA-3’) restriction enzyme. After digestion, the ends of DNA fragments were ligated with barcoded adapters using the DNA ligase. The ligated DNA was pooled and after subsequent purification, the libraries were then enriched using PCR to increase the fragment pools. The PCR products were cleaned up and checked for fragment size on a DNA analyzer. Further, the RAD-seq libraries were sequenced using an Illumina Hiseq platform.

Table 1. Planting materials and the number of samples used for ddRAD genotyping.

Planting Materials			No. of palms analyzed	Remarks
Group	Origin	No. Family		
Wild Germplasm	Angola	10	33	*Subset population* [9]
	Cameroon	6	24	*Subset population* [10]
Pure lines	Deli	8	57	Breeding lines
	AVROS	4	15	Breeding lines
	Yangambi	1	12	Breeding lines
Introgression	Compact x Ghana	5	10	
	Compact x Nigeria	1	4	
	Deli x Ekona	12	22	
	Deli x Ghana	2	9	
	Tanzania x Nigeria	3	8	
	Tanzania x Ghana	3	8	
	Tanzania x Compact	2	9	
	Yangambi x AVROS	2	8	
	AVROS-Ekona-Calabar	2	10	
Interspecific cross	O x G Hybrid	unknown	7	Commercial lines
Grand Total			236	

2.2. Allelic variation calling and downstream analysis

2.2.1. Allelic variation calling. Clean sequence reads from FASTQ files were mapped into the available reference genome of EG5 NCBI using the BWA-MEM algorithm under the Burrows-Wheeler Aligner (BWA) software [11] with a default parameter. Furthermore, under the SAM Tools package, the reads were sorted and indexed according to their genome position. The mpileup program was used for aligning the mapping results. To reduce the low-quality genotypes, only the SNPs with at least 16× sequence depth and lower than 20 % missing data were used for further analysis. Furthermore, the remaining SNPs were filtered again under criteria of the minimum of Minor Allele Frequency (MAF) of 0.05 and removed the non bi-allelic loci.

2.2.2. Allelic variant annotation and selection of effective SNPs. The variants were then annotated using the SnpEff [12] according to EG5-NCBI Annotation release ver:101 for their genomic locations. The variants were grouped into single nucleotide polymorphisms (SNPs), insertions, deletions, and multi-allelic calls. The software was also employed to predict the putative effects of mutations on gene function. SnpEff predicts the sequence ontology of the mutations, and assigned them to four predefined impact categories: *high*- (e.g. frameshift mutations and start/stop-gained variants), moderate- (e.g. missense mutations), *modifier*- (e.g. intron and intergenic mutations) and *low-putative impact* (e.g. synonymous mutations). Only the high and moderate putative impact were considered and prioritized for further SNP validations prior to designing the chip for high throughput SNP-array.

2.2.3. Genetic diversity parameters. Percentage of polymorphism across the population (% P), observed heterozygosity (Ho), expected heterozygosity (He), Shannon’s information (I) and the fixation index (F) were estimated using the Genalex v.6.5 software [13]. Principal Component Analysis of SNP collections was calculated using the GAPIT R package [14].

2.2.4. Population structure analysis and genetic clustering. Estimation of population structure was performed using the STRUCTURE ver.2.3.4 software [15] with 8 189 selected SNP loci. This software implements a model-based Bayesian clustering algorithm to correlate allele frequencies for independent runs without the need for population information. Ten independent runs were performed with different K values from 1 to 10. The Markov Chain Monte Carlo (MCMC) length of the burn-in period was set at 20 000 followed by 30 000 steps of iteration. The admixture model was implemented to obtain the optimal K value. The ΔK parameter in Structure Harvester software [16] was employed to estimate the upper-most level of structure. The genetic relationship among the individual palms was revealed by the dendrogram constructed using an unweighted Neighbor-Joining (NJ) clustering method for a dissimilarity matrix calculated with the Jaccard's coefficient. The analyses were carried out using the DARwin® v.6.0.17 software [17].

3. Result and discussion

3.1. SNP discovery from ddRAD sequencing

A total of 195.62 Gb clean data was generated from 236 palms with high quality reads prior to any filtering. The number of reads ranged from (1.23 to 32.10) × 10^6 pair-end reads per individual samples. Generally, the samples presented a mean of GC content 39.81 % ± 0.86 %. A total of 8 459 674 SNP loci was generated from reads mapping to the EG5 reference genome. SNP calling with at least 16x sequence depth, and lower than 20 % of missing data resulting in a total of 1 307 654 SNP loci. Further filtering parameters were implemented to reduce the SNPs into more effective and non-redundant markers.

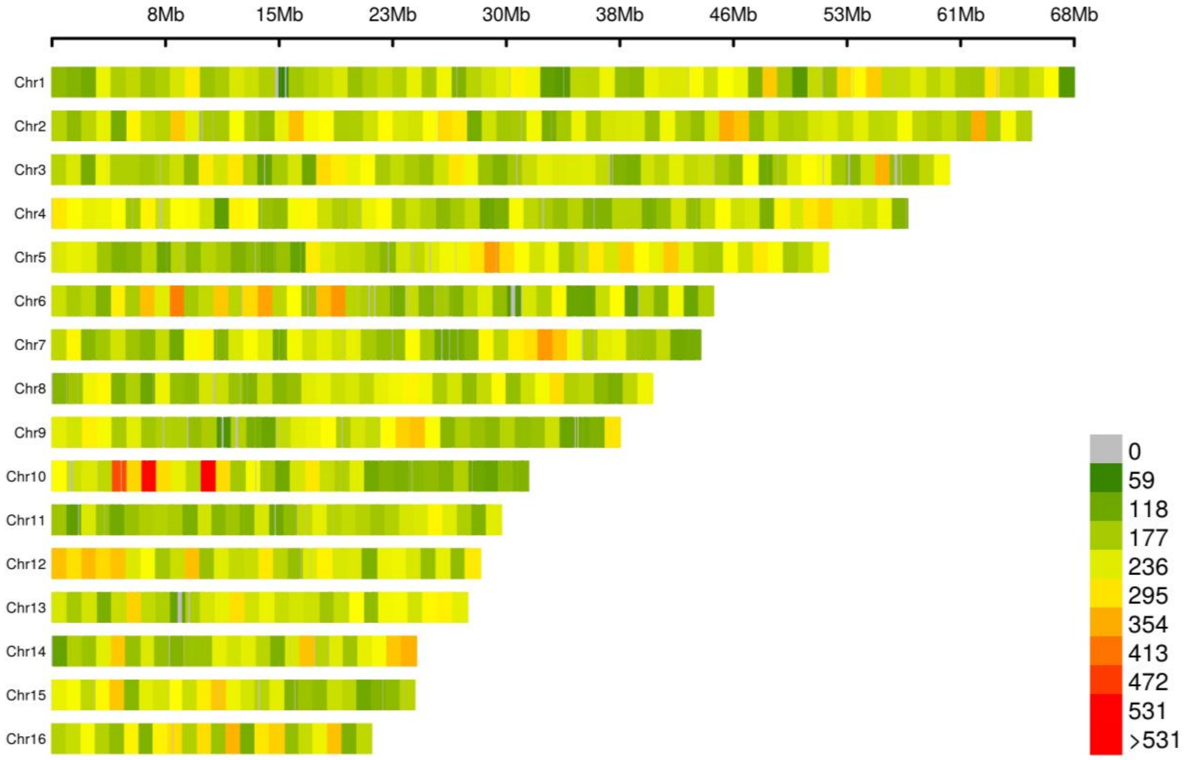

Figure 1. SNP density within 1 Mbp generated from ddRAD sequencing.

Finally, 369 200 SNP loci distributed along pseudo-chromosomes, chloroplast, and unscaffold genomes were obtained after filtering with a minimum of 5 % MAF, and only biallelic loci were considered for further analysis. SNP loci uncovered through the ddRAD sequencing were mainly distributed in non-coding portions of the genome. Overall the SNP density over the entire 16 pseudo-chromosomes was estimated to be 201 bp ± 12 bp per Mbps. There were more than 133 048 mapped

SNPs and the highest density SNPs was observed within the sixth pseudo-chromosome which reach more than 531 SNPs per Mbp (figure 1). Only 9 772 (7 %) SNP loci were located within the annotated genes (table 2) and 8 189 SNPs loci were retained after excluding the SNPs which have neighbor SNPs within 8bp at the 5'- and 3'-side windows. This SNPs collection was used for further genetic diversity and cluster analysis.

Table 2. Screening of core collection SNPs according to their location within the annotated gene and in silico putative impact

Chr	No. of biallelic SNP loci	SNP within the annotated gene	SNP without neighbor	A. High putative impact			B. Moderate putative impact		
				Splice acceptor/ Donor variants	Start/ Stop lost, Stop gained	Total of high putative impact	Missense variant	Splice region variant	Total of Moderate Putative impact
1	13 131	1 183 (9 %)	1 020	4	5	9	286	5	291
2	13 619	1 068 (8 %)	884	3	8	11	272	4	276
3	11 831	957 (8 %)	791	5	2	7	221	-	221
4	11 531	777 (7 %)	642	1	3	4	172	2	174
5	10 707	797 (7 %)	671	2	5	7	153	-	153
6	8 972	528 (6 %)	434	1	3	4	158	1	159
7	8 675	548 (6 %)	468	-	-	0	137	1	138
8	7 843	636 (8 %)	519	4	3	7	158	3	161
9	7 239	457 (6 %)	389	-	3	3	116	2	118
10	6 809	418 (6 %)	372	1	-	1	87	-	87
11	5 417	345 (6 %)	299	1	-	1	87	1	88
12	6 562	478 (7 %)	413		2	2	119	1	120
13	5 805	414 (7 %)	343	2	-	2	134	-	134
14	5 101	433 (8 %)	343	1	-	1	120	2	122
15	5 092	409 (8 %)	334	1	2	3	112	1	113
16	4 714	324 (7 %)	267	1	-	1	90	1	91
Total	133 048	9 772 (7 %)	8 189 (6 %)			63 (0.6 %)			2 446 (29 %)

In silico analysis using the SnpEff software was performed to identify the nucleotide substitutions which could alter the amino acid. Among 8 189 selected loci, only 0.26 % were giving high putative impact and 29 % of moderate putative impact mutations (the remaining loci were low or modifier impact). In the high putative impact category, equal proportion showed in both frameshift (27 SNPs; 43 %) and start/stop-gained variants (36 SNPs; 57 %) while in the moderate putative impact category, missense variant was dominated by contributing 2 422 SNP loci (99 %). Generally, over 16 pseudo-chromosomes analyzed, the Missense/Silent ratio was accounted as 1.2 132. This result was similar to a previous study in oil palm (1.2 873) as generated using the *Pst*I-*Msp*I genotyping by sequencing [18].

3.2. Genetic diversity and structure of population

In situ germplasm conservation is costly in terms of space, time and physical resources, particularly for perennial crops, such as oil palm [19]. Contrary to the breeding program requirement, the genetic resources should attempt to maintain maximum diversity so as to provide genetic variation for assembling through well-designed mating design and selection. The genetic assessment for breeding collection has a vital role in the identification of genetic resources that are most divergent from commercial breeding materials. Small effective number with genetically diverse makeup is required for conservation of germplasm.

In this study, wild germplasms were assessed using the selected sets of SNP marker in comparison with the advanced breeding materials including the inbred lines and their introgressed derivatives. Overall, a high locus polymorphism rate is shown in both wild germplasm and the inbred lines breeding materials ranging from 91 % to 99 %. In contrast, the mean of polymorphism rate of introgressed materials did not exceed 90 % except for *Deli×Ekona*, *Tanzania×Compact*, and *Avr-Eko-Cal*. progenies. It seemed that the genotype variation in the introgressed materials for particular assayed loci was low due to the domination of certain Tenera individuals compared to that derived from the crossing of two different inbred lines, giving a high heterozygosity level in the genome [8]. It was also reflected from the high level of observed heterozygosity (table 3).

Table 3. Genetic diversity parameters of various oil palm breeding populations based on 8 189 core SNP collection.

Planting Materials		N	Estimated Genetic Parameters					
Group	Origin		P	Ne	Ho	He	I	F
Wild Germplasm	Angola	33	98.8 %	1.49	0.27	0.30	0.44	0.06
	Cameroon	24	95.7 %	1.51	0.31	0.30	0.46	-0.02
Pure lines	Deli	57	99.2 %	1.49	0.29	0.30	0.45	0.03
	AVROS	15	91.7 %	1.48	0.32	0.29	0.44	-0.10
	Yangambi	12	91.6 %	1.48	0.31	0.29	0.44	-0.06
Introgression	Compact × Ghana	10	89.3 %	1.50	0.33	0.30	0.45	-0.10
	Compact × Nigeria	4	75.4 %	1.47	0.33	0.28	0.41	-0.17
	Deli × Ekona	22	97.2 %	1.51	0.32	0.31	0.47	-0.05
	Deli × Ghana	9	84.5 %	1.48	0.32	0.29	0.43	-0.11
	Tanzania × Nigeria	8	89.9 %	1.50	0.33	0.30	0.45	-0.11
	Tanzania × Ghana	8	89.2 %	1.50	0.34	0.30	0.45	-0.12
	Tanzania × Compact	9	92.5 %	1.52	0.32	0.31	0.47	-0.03
	Yangambi × AVROS	8	88.5 %	1.51	0.35	0.30	0.45	-0.15
	AVROS-Ekona-Calabar	10	93.1 %	1.52	0.34	0.31	0.47	-0.08
Interspecific cross	O × G Hybrid	7	84.6 %	1.49	0.33	0.29	0.44	-0.12
Grand Total		236						

*N: number of palms, % P: percentage of polymorphism, Ne: No. of Effective Alleles, Ho: Observed heterozygosity, He: Expected heterozygosity, I: Shannon Index, F: fixation index.

Genetic diversity between the accession groups did not much varied. Among all accessions, the lowest observed heterozygosity was found in Angola origin ($H_O = 0.270 \pm 0.002$). It was much lower compared to Cameroonian ($H_O = 0.312 \pm 0.002$) in the undomesticated population group. While, in the inbred lines of advanced breeding materials, the Deli Dura populations showed the lowest heterozygosity when compared to the AVROS and Yangambi origins. Interestingly, compared to either to the AVROS, which originated from four families or Yangambi, which was derived from a single family, the Deli Duras which were pooled from eight unrelated families had relatively higher heterozygosity. The variation within the core selected SNPs for Deli Duras was poor which may have been caused by the more intense selection for specific traits within the Deli Dura populations. This was probably the consequence of the development inbred population developed for their heterotic effects in the outcrossed Dura × Pisifera hybrids. The inbreeding effects also reflected by lower observed heterozygosity compared to expected heterozygosity under Hardy-Wienberg Equilibrium (HWE), as observed in Angola and Deli sub-population.

The fixation index (F) exhibited contrasting range from -0.17 to 0.06 with an average of -0.07. The F value represents the average deviation of the population's genotypic proportions from Hardy-Weinberg equilibrium for particular locus assayed. The negative F value represents an excess of heterozygotes. Most of the analyzed population has a negative value of this index, showing the development of heterotic population such as the D x P progenies. Values close to zero are expected under random mating, while substantial positive values indicate inbreeding or undetected null alleles. The positive value of Angola wild population can be due to the sib-matting between the palms in particular areas of sampling, while the high F index of Deli population can be due to the development of heterotic population by intercrosses within origin.

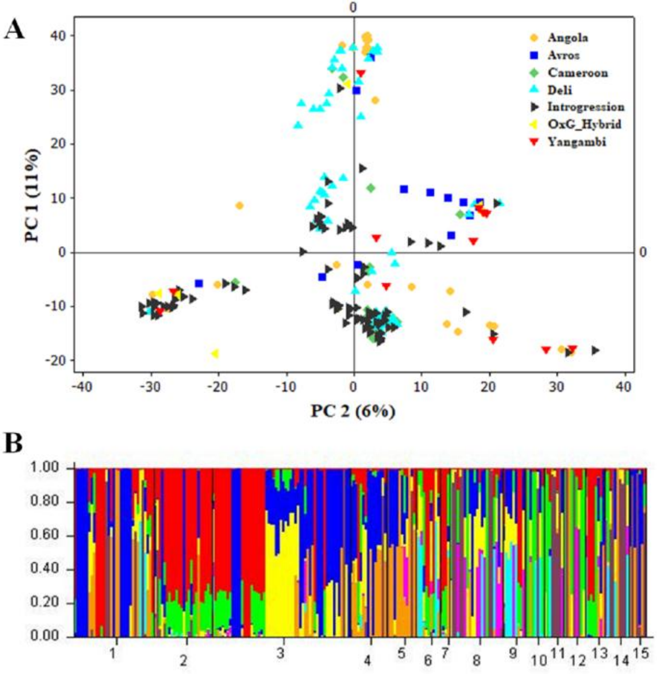

Figure 2. The relative position of individual palms projected by two Principal Component Analysis using the 8 189 SNPs collection (A). Bayesian clustering analysis of Oil palm breeding material population structure, with K = 4. Each vertical bar indicates an individual sample (B).

When the genotypic data were subjected to Principal Component Analysis (PCA) to assess the genetic diversity, the first two axes accounted for only 17 % of the total variation and could not be clearly revealed a distinct structure within the germplasm assayed (figure 2). The PCA analysis was not strong enough to stratify the group of individuals, even for OxG hybrids which sharing allelic variation from different species. The Private alleles were not in concordance with a specific genetic background. It seems that the selected SNPs loci generated from this study were the allelic variation which was commonly found in both guineensis and oleifera origins.

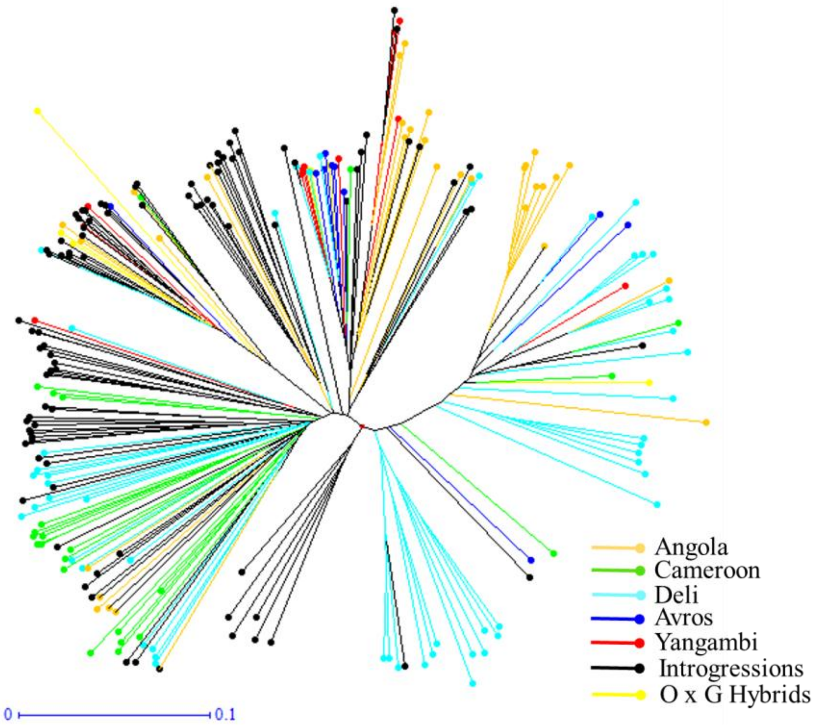

Figure 3. A Neighbor-Joining Dendogram of genetic dissimilarity among oil palm breeding collections according to the Jaccard' distance.

The genetic relationship between individuals could be represented by cluster analysis, as the variation within clusters is minimized and the variation between clusters is maximized. On one side, the unweighted Neighbor-Joining cluster analysis based on the Jaccard's genetic distance groups most of the Cameroonian palms into one main cluster sharing together with the Deli Duras and introgressed palms. On the other side, Angola population was separated into two main groups (figure 3). Similar with the PCA analysis, a clearly differentiable cluster could not be identified. The O × G hybrid also could not be exclusively separated from other genetic backgrounds.

4. Conclusion

Successful implementation of ddRAD sequencing generated 133 048 selected SNPs which distributed along the pseudo-chromosomes. Only six percent of selected SNPs were located within the annotated genes. The genetic diversity observed in germplasm collections indicate the potential sources of new allelic variations for introgression into current genetic materials which are highly inbred such as the Deli Duras. Genetic markers developed from selected SNPs can be further used for oil palm breeding programs, cultivar identification, marker-assisted selection programs, and development of the high-density map. Development of simple and cost-effective genotyping platform will be further studied for validation of selected SNPs. A high-throughput targeted-genotyping platform involving those selected SNPs in combination with with SNPs provided by the public database is being evaluated for marker-trait identification.

Acknowledgments
The authors wish to thank their respective organizations for their technical and financial support and permission to publish this paper.

References

[1] Rajanaidu N, Din A M, Marjuni M and Abdullah N 2018 Diversity in the genetic resources of oil palm *Achieving Sustainable Cultivation of Oil Palm Volume 1: Introduction, Breeding and Cultivation Techniques* 1st edition ed A Rival (Cambridge: Burleigh Dodds Science Publishing) pp 93–116
https://www.researchgate.net/publication/323215353_Diversity_in_the_genetic_resources_of_oil_palm

[2] Faizah R, Wening S, Rahmadi H Y and Purba A R 2016 Dugaan gejala depresi silang-dalam dan tingkat homozigositas populasi kelapa sawit hasil penyerbukan sendiri generasi ke-4 Sp540t dan generasi ke-5 Dura Deli [The Suspect Symptoms of Inbreeding depression and the homozygosity level of fourth generation of SP-540-T and fifth generation of Dura Deli oil palm selfing population] *J. Pen. Kelapa Sawit* **24**(2) pp 55–66 [in Bahasa Indonesia]
http://jurnalkelapasawit.iopri.org/index.php/jpks/article/view/8/7

[3] Soh A C 2018 Applications and challenges of biotechnology in oil palm breeding applications and challenges of biotechnology in oil palm breeding *IOP Conf. Ser. Earth Environ. Sci.* **183** 1–5
https://iopscience.iop.org/article/10.1088/1755-1315/183/1/012002

[4] Hayati A, Wickneswari R, Maizura I and Rajanaidu N 2004 Genetic diversity of oil palm (*Elaeis guineensis* Jacq .) germplasm collections from Africa: Implications for improvement and conservation of genetic resources *Theor. Appl. Genet.***108** 1274–84
https://www.ncbi.nlm.nih.gov/pubmed/14676949

[5] Arias D, Gonzalez M and Romero H M 2015 Genetic diversity and establishment of a core collection of oil palm (*Elaeis guineensis* Jacq.) based on molecular data. *Plant Genet. Resour.* **13**(3) 256–65
https://pubag.nal.usda.gov/catalog/5173889

[6] Peterson B K, Weber J N, Kay E H, Fisher H S and Hoekstra H E 2012 Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species *PLoS One* **7**(5) 1–11
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0037135

[7] Andrews K R, Good J R, Miller M R, Luikart G and Hohenlohe P A 2016 Harnessing the power of RADseq for ecological and evolutionary genomics *Nat. Rev. Genet.* **17**(2) 81–92
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4823021

[8] Bai B et al. 2017 Genome-wide identification of markers for selecting higher oil content in oil palm *BMC Plant Biol.* **17**(93) 1–11
https://www.ncbi.nlm.nih.gov/pubmed/28558657

[9] Sayekti U, Widyastuti U and Toruan-Mathius N 2015 Keragaman genetik kelapa sawit (*Elaeis guineensis* Jacq.) asal angola menggunakan marka SSR [Genetic diversity of the angola-originated oil palm (*Elaeis guineensis* Jacq.) using SSR Markers] *J. Agron. Indones.* **43**(2) 140–46 [in Bahasa Indonesia]
http://journal.ipb.ac.id/index.php/jurnalagronomi/article/view/10420

[10] Ajambang W, Sudarsono S, Asmono D and N Toruan 2012 Microsatellite markers reveal Cameroon’s wild oil palm population as a possible solution to broaden the genetic base in the Indonesia-Malaysia oil palm breeding programs *African J. Biotechnol.* **11**(69) 13244–49
https://academicjournals.org/journal/AJB/article-abstract/494E1FB35301

[11] Li H and Durbin R 2009 Fast and accurate short read alignment with burrows–wheeler transform *Bioinformatics* **25**(14) 1754–60
https://www.ncbi.nlm.nih.gov/pubmed/19451168

[12] Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land S J, Lu X and Ruden D M 2012 A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3 *Fly (Austin)* **6**(2) 80–92

https://www.ncbi.nlm.nih.gov/pubmed/22728672

[13] Peakall R and Smouse P E 2012 GenAlEx 6: Genetic analysis in Excel. Population genetic software for teaching and research—an update *Bioinformatics* **28**(19) 2537–39 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3463245

[14] Lipka A E, Tian F, Wang Q, Peiffer J, Li M, Bradbury P J, Gore M A, Buckler E S and Zhang Z 2012 GAPIT: Genome association and prediction integrated tool *Bioinformatics* **28**(18) 2397–99
https://www.ncbi.nlm.nih.gov/pubmed/22796960

[15] Evanno G, Regnaut S and Goudet J 2005 Detecting the number of clusters of individuals using the software structure: A simulation study *Mol. Ecol.* **14**(8) 2611–20
https://www.ncbi.nlm.nih.gov/pubmed/15969739

[16] Earl D A and Von Holdt B M 2012 Structure harvester: A website and program for visualizing structure output and implementing the Evanno method *Conserv. Genet. Resour.* **4** 359–61
https://link.springer.com/article/10.1007/s12686-011-9548-7

[17] Perrier X and Jacquemoud-Collet J P 2006 DARwin software *Montpellier*
http://darwin.cirad.fr/darwin

[18] Pootakham W, Jomchai N, Ruang-areerate P, Shearman J R, Sonthirod C, Sangsrakru D, Tragoonrung S and Tangphatsornruang S 2015 Genome-wide SNP discovery and identification of QTL associated with agronomic traits in oil palm using genotyping-by-sequencing (GBS) *Genomics* **105**(5–6) 288–95
https://www.sciencedirect.com/science/article/pii/S0888754315000361

[19] Wening S, Croxford A E, Ford C S, Thomas W T B, Forster B P, Okyere-Boateng G, Nelson S P C, Caligari P D S and Wilkinson M J 2012 Ranking the value of germplasm : new oil palm (*Elaeis guineensis*) breeding stocks as a case study *Ann. Appl. Biol.* **160** 145–56
https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1744-7348.2011.00527.x