Indonesian Journal of Agricultural Science
Vol. 17 No. 2 October 2016: 57–64
DOI: http//dx.doi.org/10.21082/ijas.v17n2.2016.p57-64

GENOMIC VARIATION OF FIVE INDONESIAN CACAO (*Theobroma cacao* L.) VARIETIES BASED ON ANALYSIS USING NEXT GENERATION SEQUENCING

*Variasi Genom Lima Varietas Kakao (*Theobroma cacao *L.) Indonesia Berdasarkan Analisis Menggunakan* Next Generation Sequencing

I Made Tasma[a], Dani Satyawan[a], Habib Rijzaani[a], Ida Rosdianti[a], Puji Lestari[a] and Rubiyo[b]

[a]*Indonesian Center for Agricultural Biotechnology and Genetic Resources Research and Development*
Jalan Tentara Pelajar No. 3A, Bogor 16111, West Java, Indonesia
Phone +62 251 8337975, Fax. +62 251 8338820
[b]*Indonesian Industrial and Beverage Crop Research Institute*
Jalan Raya Pakuwon km 2 Parungkuda, Sukabumi 43357, West Java, Indonesia
Corresponding author: imade.tasma@gmail.com

Submitted 21 April 2016; Revised 9 September 2016; Accepted 16 September 2016

ABSTRACT

Indonesian cacao productivity is still low mainly due to the lack availability of superior cacao planting materials. A new breeding method is necessary to expedite cacao yield improvement programs. To date, no study has yet been done to characterize Indonesian cacao varieties at the whole genome level. The objective of this study was to characterize genomic variation of five superior Indonesian cacao varieties using next-generation sequencing. Genetic materials used were five Indonesian cacao varieties, i.e. ICCRI2, ICCRI3, ICCRI4, SUL2 and ICS13. Genome sequences were mapped to the cacao reference genome sequence of Criollo variety. Sequence alignment and genomic variation discovery were done using Bowtie2 and mpileup software of Samtools, respectively. A total of 2,326,088 single nucleotide polymorphisms (SNPs) and 362,081 insertions and deletions (Indels) were obtained from this study. In average, a DNA variant was identified in every 121 nucleotides of the genome sequence. Most of the DNA variants were located outside the genes. Only 347,907 SNPs and Indels (13.18%) were located within protein coding region (exon). Among the DNA variations within exon, 188,949 SNPs caused missense mutation and 1,535 SNPs induced nonsense mutation. Unique gene-based SNPs were also discovered from this study that can be used as fingerprints for the particular cacao variety. The DNA variants obtained were excellent DNA marker resources to support cacao breeding programs. The SNPs discovered are useful as materials for genome-wide SNP chip development to be used for gene and QTL tagging of important traits for expediting national cacao breeding program.

[***Keywords*****:** *Theobroma cacao*, genome sequencing, genome variation, SNP, next generation sequencing].

ABSTRAK

Produktivitas kakao Indonesia masih rendah antara lain karena kurang tersedianya bahan tanaman unggul. Metode pemuliaan baru perlu diterapkan untuk mempercepat program pemuliaan kakao nasional. Sampai saat ini belum ada penelitian untuk mengkarakterisasi varietas kakao Indonesia pada level genom total. Penelitian ini bertujuan untuk mengidentifikasi variasi genom varietas unggul kakao Indonesia menggunakan next generation sequencing*. Materi genetik yang digunakan adalah lima varietas unggul kakao Indonesia, yaitu ICCRI2, ICCRI3, ICCRI4, SUL2, dan ICS13. Data sekuen genom kelima varietas tersebut dijajarkan dengan sekuen genom rujukan kakao varietas Criollo. Penjajaran sekuen dilakukan menggunakan* software *Bowtie2 dan identifikasi variasi genom (SNP dan Indel) dilakukan dengan* software *mpileup dari Samtools. Penelitian ini menghasilkan variasi genom sebanyak 2.688.169 yang terdiri atas 2.326.088 SNP dan 362.081 insersi dan delesi (Indel). Secara rata-rata satu variasi genom (SNP atau Indel) ditemukan pada setiap 121 basa dari sekuen genom kakao. Dari seluruh SNP yang diidentifikasi, 347.907 SNP (13,18%) berlokasi pada* protein coding region*. Dari jumlah ini, 188.949 SNP menyebabkan mutasi yang mengubah susunan asam amino pada protein (*missense mutation*) dan 1.535 SNP menyebabkan mutasi yang menghasilkan* stop codon *(*nonsense mutation*). Ditemukan juga SNP berbasis gen yang unik pada setiap genotipe kakao yang dapat digunakan sebagai sidik jari dari setiap genotipe kakao yang diuji. Variasi genom yang dihasilkan merupakan sumber daya marka DNA bernilai tinggi untuk studi genetika dan pemuliaan kakao. SNP hasil penelitian ini dapat digunakan sebagai materi untuk pembuatan* SNP chip *kapasitas tinggi yang bermanfaat untuk pelabelan gen unggul dan QTL yang terkait karakter penting untuk mendukung percepatan program pemuliaan kakao nasional.*

[***Kata kunci*****:** *Theobroma cacao*, sekuensing genom total, variasi genom, SNP, *next generation sequencing*]

INTRODUCTION

Successful plant breeding programs depend on the accurate characterization of plant genetic resources (PGR) for gene identification, breeding and propagation. Originally, cacao PGR collections were curated and characterized based on morphological and agronomic characteristics of individual clones

(Engels 1983; Iwaro and Butler 2000). Efforts have been done to assess, characterize and utilize cacao PGR collections based on genetic diversity, population structure and evolutionary relationships using molecular markers. The advanced genomic technologies have also been applied such as next generation sequencing and high throughput genotyping systems (Motamayor *et al*. 2008; Irish *et al*. 2010; Boza *et al*. 2013; Tasma 2015a, 2015b; Tasma *et al.* 2015). Other applications of molecular markers in integrated cacao genetics and breeding programs are addressed to use them to obtain improved varieties and maintain these improved varieties in breeding programs (Risterucci *et al.* 2000; Irish *et al*. 2010).

Indonesia is one of the top 10 cacao-producing countries and is the third most cacao producer in the world (Rubiyo *et al*. 2015). Nevertheless, its cacao productivity remains low with an average national yield of 740,510 tons (BPS 2016). The main constraints are disease and insect attacks such as fruit rot caused by *Phythophthora* sp., vascular streak dieback (VSD) caused by *Ceratobasidium theobromae* and cocoa pod borer by *Conopomorpha cramerella* (Wardojo 1992; Susilo *et al*. 2011). Another problem is the limited information on genetics and genomics of cacao. Therefore, resequencing of the genomes of Indonesian cacao varieties using NGS technology Hiseq is important to accelerate cacao breeding program and understand the genetic and genomic information on cacao.

Several types of molecular markers have been applied in cacao improvement programs, such as RAPD, RFLP, AFLP, SSR and SNP (Risterucci *et al.* 2000; Schnell *et al*. 2005; Irish *et al.* 2010). Original genetic map of cacao genomes was established using these markers. The initial cacao genetic map consisted of >250 SSRs and >400 RFLP, RAPD, AFLP and isozyme markers, covering ~900 cM of the 10 cacao chromosomes with an average marker distance of ~2 cM (Risterucci *et al.* 2000; Pugh *et al*. 2004). Among these DNA markers used, simple sequence repeats (SSRs) markers have successfully been used and have served as genetic markers in most cacao studies since ten years ago (Irish *et al.* 2010; Livingstone *et al.* 2011; Boza *et al*. 2013). However, single-nucleotide polymorphism (SNP), the most abundant, high throughput and reliable marker, is rapidly overtaking SSRs (Tasma 2014; Thomson 2014). The use of SNPs has revolutionized genetic research on many species, including cacao (Tasma 2015a, 2016a). Thousands SNPs have been discovered in cacao and used for comparative genomic studies, production of consensus genetic maps, marker-assisted breeding and curation of germplasm collections (Schnell *et al*. 2005; Argout *et al.* 2011; Allegre *et al*. 2012; Tasma 2015a).

The cacao genome reference sequence has been completed and is available for public uses (Argout *et al.* 2011). The sequence was based on a Criollo variety. From the reference sequence, cacao genome contained a total of 28,798 genes (Argout *et al.* 2011). The availability of the reference sequence opens the door for cacao scientists to exploit cacao germplasm collections into a more comprehensive, more effective and more efficient manner for breeding programs to identify genes and quantitative trait loci (QTLs). The reference genome sequence is also important for dissecting for highly economic traits such as yield components, disease and insect resistance, cocoa bean quality and nutrition related traits (Tasma 2015a, 2015b). In addition to the reference genome, Matina 1-6 which was sequenced by Motamayor *et al*. (2013) has substantially increased the number of genetic resources for identification of novel molecular markers. In parallele with genomics, bioinformatic technology becomes the key in identifying the important and useful genes within billions bases of genome sequences (Tasma 2015b, 2016b) for cacao improvement program.

Based on the genome characteristic, cacao is a diploid crop with 20 pairs of chromosomes (2n = 2x = 20). The genome size of cacao is relatively small (430 Mb), similar to the genome size of rice. In term of genome size, cacao may be considered as an estate crop model due to the less complexity in handling the genome compared to other estate crops commonly cultivated in the tropics that generally have much bigger genome sizes (e.g. oil palm of 1,800 Mb and sugarcane of 16,000 Mb) (Argout *et al*. 2011; Tasma 2016b) .

Since one of the efforts to improve Indonesian cacao productivity is by improving plant genetic materials, genomic technology is expected to assist breeding programs. Whole genome resequencing of various cacao germplasm accessions is required for effective identification of economically important genes and quantitative characters (QTLs) to be used in cacao improvement program. The genes together with the marker-based selection technology will improve the accuracy and decrease perennial crop breeding cycles that will expidite selection process (Schnell *et al*. 2005; Tasma 2015a). In addition, resequencing studies will provide better information on the genomics and genetics of the available cacao germplasm collections.

Breeding technology based on genomic data through marker-assisted selection (MAS) and marker-aided backcrossing (MAB) will also be useful for decreasing breeding cycle to be a half of that used in the classical cacao breeding programs (Thomson 2014; Tasma 2015b). This approach could make breeding programs be more efficient and effective as the trait selection can be done at the early stages of plant growth without waiting the particular plant growth stage when the particular traits of interests are expressed (Varsney *et al*. 2009; Tasma 2015b).

Next generation sequencing (NGS) has capabilities to sequence whole complexed genomes in a shorter time with lower sequencing cost (Schuster 2008; Pattersson *et al*. 2009; Zhang *et al*. 2011). This technology opens the way to resequence each accession of cacao collections to comprehensively understand genetic make-up of each cacao accession through resequencing the genome simultaneously. By resequencing, discovery of genetic variations of the targeted traits and identification of plants having the genes and alleles to be used in a breeding program of that particular trait would be achieved (Tasma 2016a).

The objectives of this study were to identify genomic variations based on the whole genome sequences of five Indonesian cacao varieties compared to that of the cacao reference genome sequence of Criollo variety. The nucleotide variations in the genes would be of interest to identify variations affecting the protein in coding regions from where the gene of interest could be isolated.

MATERIALS AND METHODS

Genetic Materials

Five Indonesian superior cacao varieties (ICCRI2, ICCRI3, ICCRI4, SUL2 and ICS13) belonged to the Indonesian Research Institute for Industrial and Beverage Crops (RISTRI), Pakuwon, Sukabumi, West Java were used in this study. The varieties have been grown in the RISTRI cacao field collections.

Isolation of Genomic DNA

Genomic DNA was isolated from healty young leaf taken from the first fully open leaf from the shoot tip of each cacao variety by using cetyl trimethyl ammonium bromide (CTAB) following the method of Michiels *et al*. (2003) with minor modification. This protocol was designed specifically for isolating DNA from leaf of the latex-rich plant as previously reported (Satyawan dan Tasma 2011a, 2011b). The isolated DNA was diluted in 50 µL TE buffer (10 mM Tris pH 8, 1 mM EDTA pH 7.5). DNA concentration and purity was then measured using a nanodrop spectro-photometer (Thermo Scientific, USA). DNA concentration was high ranged from 198.1 ng μL^{-1} (ICCRI02) to 303.7 ng μL^{-1} (SUL02). DNA quality was also high with A260/280 ratio of 1.992-2.0 (Tasma *et al*. 2012). DNA was also electrophoresed in 1% agarose gel (Sambrook *et al*. 1989), and the DNA bands were visualized with a Chemidoc (BioRad, California, USA). The DNA bands were bright and intact indicating the high quality and high concentration of the isolated DNA (Tasma *et al*. 2012).

Construction and Validation of Genomic DNA Libraries

Cacao genomic library construction was conducted using the Illumina TruSeq DNA low throughput (LT) protocol (Illumina Inc., USA) using a recommended procedure (Tasma *et al*. 2012). In brief, the procedure included DNA fragmentation, DNA end modification, adenylation of 3'end, adaptor ligation and purification, PCR amplification, and validation of the constructed cacao genomic library. The library concentration was measured using RT PCR method.

Whole Genome DNA Sequencing of Cacao Genomic Libraries

Genome sequencing of cacao genomic libraries consisted of two steps, i.e. DNA clusterization in a DNA cluster platform (cBot) and sequencing of the clustered genomic DNA using the NGS HiSeq2000. DNA cluster generation was conducted using the cluster generation protocol, reagents and kits from the manufacturer (Illumina Inc., CA, USA; Quail 2008). A more detailed protocol on DNA cluster generation was previously reported (Tasma *et al*. 2012). Sequencing of the clustered libraries was performed using sequencing reagents and kits from Illumina (Illumina Inc., California, USA) according to the sequencing procedure as previously reported by Tasma *et al*. (2012). A paired-end genome deep sequencing method was accomplished for the five cacao varieties using a total of 200 cycles for both sequencing reads.

Analysis of Sequence Data for Genome Variation Characterization

Resequence data of the five cacao varieties were bioinformatically analyzed by aligning the sequences with that of cacao genome reference derived from Criollo variety (Argout *et al.* 2011) using Bowtie2 software (Langmead and Salzberg 2012). Based on the sequencing alignment, genome variation was characterized using a computer software *mpileup* within Samtools (Ewing and Green 1998; Li *et al*. 2009). Anotation of the location and prediction of the effect of the identified SNPs/Indels were conducted using snpEff software (Cingolani *et al*. 2012). Finally, a genetic diversity analysis of the five genome sequences was conducted using DarWin software (Perrier and Jacquemoud-Collet 2006).

RESULTS AND DISCUSSION

Genome Sequencing Coverage

The average of sequencing coverage of the NGS was 30, meaning that in average the five superior cacao varieties were sequenced 30 times (Fig. 1). The average depth exceeding 30 times was the de facto standard (Ahn *et al*. 2009) and the type of genome sequence coverage was classified as a deep sequencing (Sims *et al*. 2014). The coverage obtained from this study was very excellent to obtain high quality SNPs and Indels as the more sequence overlapping would be obtained within the identified SNP or Indel that increase SNP/Indel discovery accuracy. The identified SNPs/Indels, therefore, will be more reliable for further analysis and for breeding application purposes.

Number and Location of Genome Variations

Alignment of the five genome sequences with the sequence of cacao reference genome of Criollo variety (Argout *et al*. 2011) resulted a total of 2,688,169 variations consisting of 2,326,088 SNPs and 362,081 Indels. This number was identified across chromosomal regions of 327,353,121 bp obtaining the change rate of 121 (Table 1). This means that a DNA variation (SNP or Indel) existed in every 121 nucleotides of the cacao genome sequence. These nucleotide frequency changes are slightly higher than that observed in soybean in which one DNA variation was detected within every 308 bases (Satyawan *et al*. 2014). The differences were mostly due to the different modes of reproduction of the two crops in which cacao is a cross pollinated crop, while soybean is a self-pollinated one, making more frequent DNA variations discovered in cacao genome than that in soybean genome.

The DNA variation changes observed were different across different cacao genome regions. Most of the genome variations were discovered outside the protein coding region (exon), including the upstream region (38.61%), downstream region (37.01%), intergenic region (4.23%) and intron (9.54%). DNA variations located within the gene sequences (exon, intron and untranslated regions)

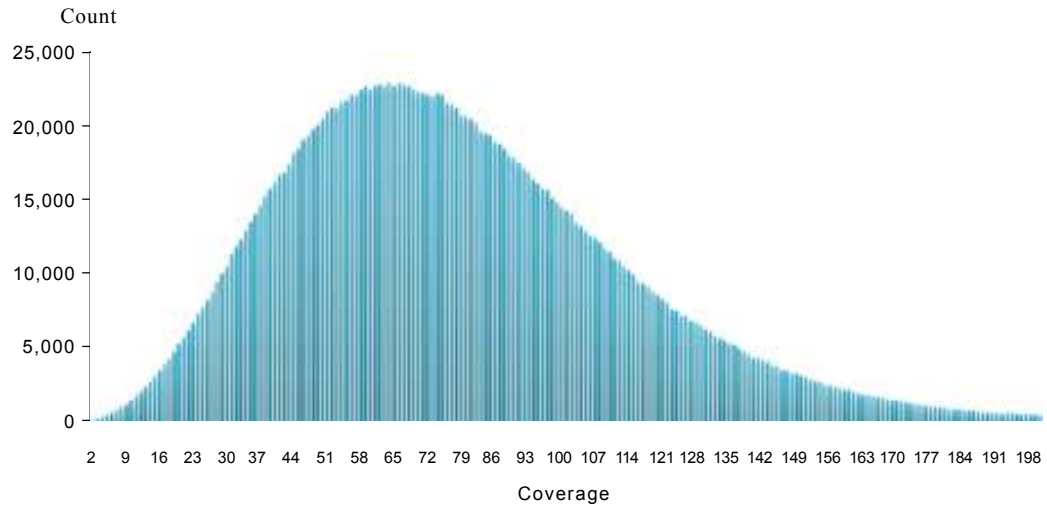

Fig. 1. Genomic sequence coverage of five Indonesian cacao varieties. The paired-end sequencing (using 200 sequencing cycles) was done in the NGS sequencing platform of HiSeq2000.

Table 1. The number of DNA changes (DNA variations) and the rate of DNA changes detected in each of the cacao chromosomes.

Chromosome	Chromosome length (bp)	DNA variation changes	Change rate
Tc00	108,886,888	835,679	130
Tc01	31,268,538	254,505	122
Tc02	27,754,001	232,265	119
Tc03	25,475,297	198,549	128
Tc04	23,504,306	166,738	140
Tc05	25,651,337	208,396	123
Tc06	15,484,475	138,079	112
Tc07	14,169,093	106,745	132
Tc08	11,535,834	111,952	103
Tc09	28,459,094	284,016	100
Tc10	15,164,258	151,245	100
Total/average	327,353,121	2,688,169	121

were 20.16%. Among those, only 13.1% were located within protein coding region (exon). Most of the genome variations (79.84%) from this study, therefore, were outside the gene regions. This study result is similar to that identified in soybean genome sequences in which 88.19% of the total genome variations were located outside the gene regions (Satyawan *et al*. 2014; Tasma *et al.* 2015). However, genome variations observed within gene regions were much higher in this study (20.16%) compared to that in soybean (11.81%).

The number of variations within exon was also higher in this study (13.18%) compared to that observed in soybean (2.16%) as reported by Satyawan *et al*. (2014). This is most likely due to the difference in genome sizes and genome complexities between the two crop species. The size of soybean genome is approximately three times of the size of cacao genome. Soybean genome is also much richer with repeated and duplicated regions compared to that of cacao genome (Schmutz *et al*. 2010; Argout *et al*. 2011). In addition, it is well known that the number of genes across plant species is relatively similar regardless of their genome sizes. For that reason, a much higher percentage of genome variations was observed in cacao genic regions than that in the soybean genic regions.

Types of Genome Variations

Among the total DNA variations observed, 2,326,088 were SNPs and 362,081 were Indels (Table 2). For breeding and gene discovery purposes, we are more interested to further characterize DNA variations located within the genic regions (genic SNPs and genic Indels). Out of 857,506 genic SNPs identified from this study, 347,907 SNPs were located within the exon (Table 2). Of those, 108,720 SNPs were synonymous (SNPs that do not change amino acid composition of the encoded proteins and do not affect gene function) and 214,068 SNPs were nonsynonymous (SNPs that change amino acid composition of the proteins encoded by the genes containing the SNPs). This later type of SNPs certainly affects protein composition and gene function. Most of the nonsynonymous SNPs (188,949) were missense mutations. Other SNPs caused either changes in stop and start codons or splice sites (Table 2) during RNA processing to obtain mature RNA for translation. These latest SNPs types also significantly affect gene function through the loss of functions of the genes containing such types of SNPs. A low number of genic Indels were also observed totaling of 219,126. Among those, only 6,299 Indels were located within exons. The remaining Indels were located within introns that were spliced during RNA processing. Most of the exonic Indels (5,359) caused protein frameships. Proteins encoded by the genes phenotypically would be non-functional causing the loss of functions of such particular SNP

Table 2. Number and types of DNA sequence variations detected based on the sequence data of five Indonesian cacao varieties.

Types of DNA variations	Number of DNA variations
Number of total variants	
SNP	2,326,088
Indel	362,081
Genic SNPs	
In introns	509,599
In exon	347,907
Synonymous	108,720
Nonsynonymous	214,068
Missense	188,949
Start codons gained	8,133
Start codons lost	616
Stop codons gained	10,078
Stop codons lost	1,355
Splice site acceptor	2,536
Splice site donor	2,401
Genic Indels	
In introns	212,827
In exon	6,299
Frameshift	5,359
Codon change plus codon deletion	173
Codon change plus codon insertion	304
Codon deletion	260
Codon insertion	203

containing genes. The remaining minor indels caused codon changes (Table 2) that could affect or not affect gene functions as some codons could be determined by different types of nucleotide combinations.

The SNPs and Indels that affect gene function became our SNP type of interest for gene discovery purposes mainly genes encoding the agronomically and economically important traits. Such genes included those affecting yield, environmentally adapted genes (e.g. biotic and abiotic stress genes), and adaptation genes including tropically adapted genes for crops originated from subtropical areas such as soybean and wheat, quality traits, and the ones for coping global warming phenomenon.

Common and Unique SNP Loci in Five Cacao Varieties

Taken SNP samples located within the exon and the DNA fragments containing the SNPs were sequenced at least five times within each of the cacao variety studied. Among the 108,720 SNPs located in exons, 2,451 SNPs met the criteria. Each cacao variety was then analyzed and scored using the selected 2,451 SNP loci.

The results showed that five cacao varieties demonstrated common DNA variations in a total of 854 SNP loci that were different from that of the cacao genome reference sequence of Criollo variety (Table 3). Hundreds of unique SNP loci were found within a particular cacao variety. A total of 234 unique SNP loci were observed in ICCRI 02; 315 unique SNP loci were found in ICCRI 03; 202 loci were discovered in ICCRI 04; 452 unique loci were observed in Sulawesi 2, and 394 unique SNP loci were found in the genome of ICS13 (Table 3). These unique SNP loci can be used as fingerprints for that specific variety and as a unique identity of the cacao variety. Such kind of DNA fingerprints can be used for protecting superior germplasm from unresponsible use of the genetic materials (Tasma 2015a, 2016a). The SNP loci, therefore, can be used for monitoring the utilization of the genetic materials with high precision in a fast assay manner.

We assessed the genetic relatedness of the five cacao varieties compared to that of Criollo variety, the reference genom, using mutation data. Three varieties (ICCRI 04, ICCRI 02 and ICCRI 03) grouped separately from the other two varieties (Sulawesi12 and ACS13). ICCRI 04 and ICCRI 02 were closer each other compared to ICCRI03. As expected, there was a far distance between bulk cocoa (Sulawesi12 and ACS13) and edel/fine cocoa (ICCRI02, ICCRI03 and ICCRI04), reflecting the distinctive of their genome which supports quality phenotypes of either bulk or fine cacao type. It is notable that Criollo, known as producing edel clones (Susilo *et al*. 2011), was closer to group of bulk cacao than that of edel/fine cacao in this study (Fig. 2). Clearly, a greater variation existed between cacao groups, in contrast to close relationship of varieties/genotypes within the same group. Incompatible crossing between the groups of edel/fine cacao and bulk cacao is also likely influenced by their differences in the genome. Therefore, this genetic relatedness information could be very useful as a basis for parental screening in cacao breeding improvement programs.

Table 3. Common and unique SNP markers found in five Indonesian cacao varieties.

Genotype	Number of unique and common SNPs
Unique genic SNP[a]	
ICCRI 02	234
ICCRI 03	315
ICRRI 04	202
SUL 2	452
ICS 13	394
Common genic SNP[b] (found in all five cacao genotypes)	854
Total SNPs	2,451

[a]Unique genic SNP is a SNP derived from exon that is unique to a particular cacao genotype analyzed in this study (i.e. ICCRI 02, ICCRI 03, ICRRI 04, SUL 2 or ICS 13).

[b]Common genic SNP is a SNP derived from exon that belonged to all cacao genotypes analyzed in this study.

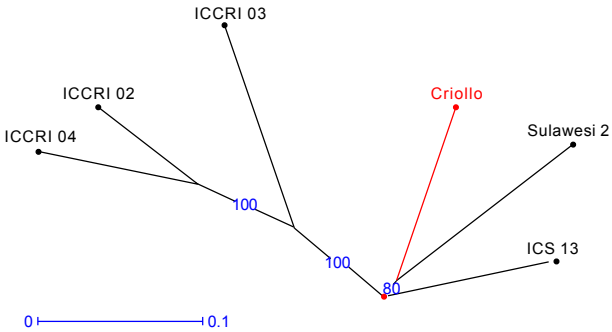

Fig. 2. Phylogenetic relatedness among the five superior Indonesian cacao varieties in comparison with that of cultivar Criollo fom where the cacao reference genome sequence was derived. Variety labeled with red color denotes the reference genome cultivar, and those labeled with blue color are the Indonesian cacao varieties used in this study.

CONCLUSION

Alignment of the resequence data derived from five superior Indonesian cacao varieties with the cacao reference genome sequence of Criollo variety resulted a total of 2.6 million DNA variations consisted of 2.3 million SNPs and 0.3 million Indels. In average, one DNA variation was obtained per 121 nucleotides of the cacao genome sequences. Most of the DNA variations were located outside genic regions (79.842%) and 20.158% were within genic region. Among those, only 13.18% were located within protein coding region (exon). Phylogenetic analysis supports the distinctive genome variation between edel/fine and bulk cacao groups. The DNA variation obtained from this study is very useful for high throughput DNA marker development to expedite national cacao breeding programs.

ACKNOWLEDGEMENT

This study was funded by the 2011 and 2012 Fiscal Year Governmental Funding Under ICABIOGRAD-IAARD. The authors thank Mr. Andi Kosasih for his involvement in isolating cacao genomic DNA used in this study.

REFERENCES

Ahn, S.M., T.H. Kim, S. Lee, D. Kim, H. Ghang, B.C. Kim, S.Y. Kim, W.Y. Kim, D. Park, Y.S. Lee, S. Kim, R. Reja, S. Jho, C.G. Kim, J.Y. Cha, K.H. Kim, B. Lee, J. Bhak and S.J. Kim. 2009. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res. 19: 1622–1629.

Allegre, M., X. Argout, M. Boccara, O. Fouet, Y. Roguet, A. Berard, J.M. Thevenin, A. Chauveau, R. Rivallan, D. Clement, B. Courtois, K. Gramacho, A. Boland-Auge, M. Tahi, P. Umaharan, D. Brunel and C. Lanaud. 2012. Discovery and mapping of a new expressed sequence tag-single nucleotide polymorphism and simple sequence repeat panel for large-scale genetic studies and breeding of *Theobroma cacao* L. DNA Res. 19(1): 23–25.

Argout, X., J. Salse, J.-M. Aury, M.J. Guiltinan, G. Droc, J. Gouzy, M. Allegre, C. Chaparro, T. Legavre1, S.N. Maximova, M. Abrouk, F. Murat, O. Fouet, J. Poulain, M. Ruiz, Y. Roguet, M. Rodier-Goud, J.F. Barbosa-Neto, F. Sabot, D. Kudrna, J.S.S. Ammiraju, S.C. Schuster, J.E. Carlson, E. Sallet, T. Schiex, A. Dievart, M. Kramer, L. Gelley, Z. Shi, A. Bérard, C. Viot, M. Boccara, A.M. Risterucci, V. Guignon, X. Sabau and M.J. Axtell. 2011. The genome of *Theobroma cacao*. Nature Genetics 4(2): 101–108.

Boza, E.J., B.M. Irish, A.W. Meerow, C.L. Tondo, O.A. Rodrý´guez, M. Ventura-Lo´pez, J.A. Go´mez, J.M. Moore, D. Zhang, J.C. Motamayor and R.J. Schnell. 2013. Genetic diversity, conservation,and utilization of *Theobroma cacao* L.: genetic resources in the Dominican Republic. Genet. Resour. Crop Evol. 60: 605–619.

BPS. 2016. Badan Pusat Statistik, Indonesia. https://www.bps.go.id. [April 17, 2016)].

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

Engels, J. 1983. A systematic description of cacao clones. III. Relationships between clones, between characteristics and some consequences for the cacao breeding. Euphytica 32: 719–33.

Ewing, B. and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8: 186–194.

Irish, B.M., R. Goenaga, D. Zhang, R. Schnell, J. S. Brown and J.C. Motamayor. 2010. Microsatellite fingerprinting of the USDA-ARS Tropical Agriculture Research Station Cacao (*Theobroma cacao* L.) germplasm collection. Crop Sci. 50: 656–567.

Iwaro, A.D. and D.P. Butler. 2000. Germplasm enhancement for resistance to black pod and witches' broom diseases. *In*: Proceedings of the 13th International Cocoa Research Conference: Towards the Effective and Optimum Promotion of Cocoa through Research and Development. Cocoa Producers Alliance, Lagos, Nigeria.

Langmead, B and S.L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9(4): 357–359.

Li, H., B. Handsaker and A. Wysoker. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–2079.

Livingstone, D., J.C. Motamayor, R.J. Schnell, K. Cariaga, B. Freeman, A. Meerow, J. Brown and D.N. Kuhn. 2011. Development of single nucleotide polymorphism markers in *Theobroma cacao* and comparison to simple sequence repeat markers for genotyping of Cameroon clones. Mol. Breed. 27: 93–106.

Michiels, A.N., E.W. van Den, M. Tucker, R. Liesbet and L. van Andre. 2003. Extraction of high quality genomic DNA from latex containing plants. Anal. Biochem. 315: 85–89.

Motamayor, J.C., P. Lachenaud, J.W. da Silva Mota, R. Loor, D.N. Kuhn, J.S. Brown, and R.J. Schnell. 2008. Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L.). PLoS ONE: 3: e3311.

Motamayor, J.C., K. Mockaitis, J. Schmutz, N. Haiminen, D. Livingstone, O. Cornejo, S. D. Findley, P. Zheng, F. Utro, S. Royaert, C. Saski, J. Jenkins, R. Podicheti, M. Zhao, B. E. Scheffler, J.C. Stack, F.A. Feltus, G.M. Mustiga, F. Amores, W. Phillips, J.P. Marelli, G.D. May; H. Shapiro, J. Ma, C.D. Bustamante, R.J. Schnell, D. Main, D. Gilbert, L. Parida and D.N. Kuhn. 2013. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. Genome Biol. 14: R53.

Pattersson, E., J. Lundeberg, and A. Ahmadian. 2009. Generations of sequencing technologies. Genomics 93: 105–111.

Perrier, X. and J.P. Jacquemoud-Collet. 2006. DARwin Software. http:// //darwin.cirad.fr/darwin

Pugh, T., O. Fouet, A.M. Risterucci, P. Brottier, M. Abouladze, C. Deletrez, B. Courtois, D. Clement, P. Larmande and J.A.K. Ngoran. 2004. A new cacao linkage map based on codominant markers: development and integration of 201 new microsatellite markers. Theor. Appl. Genet. 108: 1151–1161.

Quail, S. 2008. A large genome center's improvements to the illumina sequencing system. Nature Methods 5: 1005–1010.

Risterucci, A.M., L. Grivet, J. A. K. N'Goran, I. Pieretti, M. H. Flament, and C. Lanaud. 2000. A high-density linkage map of *Theobroma cacao* L. Theor. Appl. Genet. 101: 948–955.

Rubiyo, N.K. Izzah, I. Sulistiyorini and C. Tresniawati. 2015. Evaluation of genetic diversity in cacao collected from Kolaka, South Sulawesi, using SSR markers. Indones. J. Agric. Sci. 16(2): 71–78.

Sambrook, J., E.F. Fritch and T. Maniatis. 1989. Molecular Cloning: A laboratory Manual. Cold Spring Harbor Laboratory Press.

Satyawan, D. and I.M. Tasma. 2011a. Genetic diversity analysis of *Jatropha curcas* provenances assessed with randomly amplified polymorphic DNA markers. J. AgroBiogen 7(1): 47–55.

Satyawan, D. and I.M. Tasma. 2011b . DNA markers applicable for genetic mapping of *J. curcas* genome. Bulettin Riset Tanaman Industri 2 (3): 411-419.

Satyawan, D., H. Rijzaani and I.M. Tasma. 2014. Characterization of genomic variation in Indonesian soybean (*Glycine max*) using next-generation sequencing. Plant Genet. Resour. S1: S109–S113.

Schnell, R.J., C. T. Olano, J.S. Brown, A.W. Meerow, C. Cerventes-Martinez, C. Nagai and J.C. Motamayor. 2005. Retrospective determination of the parental population of superior cacao (*Theobroma cacao* L.) seedlings and association of microsatellite alleles with productivity. J. Am. Soc. Hortic. Sci. 130: 181–190.

Schmutz, J.S.B., J. Schlueter, J. Ma, T. Mitros, W. Nelson, D.L. Hyten, Q. Song, J.J. Thelen, J. Cheng, D. Xu, U. Hellsten, G.D. May, Y. Yu, T. Sakurai, T. Umesawa, M.K. Bhttacahrrya, D. Sandhy, B. Valliyodan, E. Linquist, M. Peto, D. Grant, S. Shu, D. Goodstein, R.C. Shoemaker and S.A. Jackson *et al.* (44 authors). 2010. Genome sequence of the paleopolyploid soybean. Nature 463: 178–183.

Schuster, S.C. 2008. Next generation sequencing transform today's biology. Nat. Method 5: 16–18.

Sims, D., I. Sudbery, N.E. Llott, A. Heger and C.P. Ponting. 2014. Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. 15: 121–132.

Susilo, A.W., D. Zhang, L.A. Motilal, S. Mischke and L.W. Meinhardt. 2011. Assessing genetic diversity in Java fine-flavor cocoa (*Theobroma cacao* L.) germplasm by using simple sequence repeat (SSR) markers. Trop. Agric. Dev. 55(2): 84–92.

Tasma, I.M., D. Satyawan, H. Rijzaani, dan Rubiyo. 2012. Konstruksi pustaka genom kakao (*Theobroma cacao* L.) untuk sekuensing genom total menggunakan *next generation sequencing* HiSeq2000. Buletin Riset Tanaman Rempah dan Aneka Tanaman Industri 3(2): 99–108.

Tasma, I.M. 2014. *Single nucleotide polymorphism* (SNP) sebagai marka DNA masa depan. Warta Biogen 10(3): 7–10.

Tasma, I.M. 2015a. Pemanfaatan teknologi sekuensing genom untuk mempercepat program pemuliaan tanaman. J. Litbang Pert. 34(4): 159–168.

Tasma, I.M. 2015b. The use of advanced genomic platforms to accelerate breeding programs of the Indonesian Agency for Agricultural Research and Development. Internat. J. Biosci. Biotechnol. 2(2): 43–53.

Tasma, I.M., H. Rijzaani, D. Satyawan, P. Lestari, D.W. Utami, I. Rosdianti, A.R. Purba, E. Mansyah, A. Sutanto, R. Kirana, Kusmana, A. Anggraeni, M. Pabendon and Rubiyo. 2015. Next-gen-based DNA marker development of several importance crop and animal species. Manuscript Presented at the 13th SABRAO Congress and International Conference, 14–16 September 2015, IPB International Convention Center, Bogor, Indonesia. 8 pp.

Tasma, I.M. 2016a. Resekuensing genom, metode baru karakterisasi variasi SDG tanaman secara komprehensif mendukung akselerasi program pemuliaan tanaman. Warta Biogen 12(1): 2–6.

Tasma, I.M. 2016b. Pemanfaatan teknologi genomika dan transformasi genetik untuk meningkatkan produktivitas kelapa sawit. Perspektif 15(1): 51–73.

Thomson, M.J. 2014. High-throughput SNP genotyping to accelerate crop improvement. Plant Breed. Biotechnol. 2(3): 195–212.

Varshney, R.K., S.N. Nayak, G.D. May, and J.A. Jackson. 2009. Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends in Biotechnol. 9: 522–530.

Wardojo, S. 1992. Major pest and dieases of cocoa in Indonesia. *In* P.J. Keane and C.A.J. Petter (Eds.). Cocoa Pest and Disease Management In Southeast Asia and Australia. FAO Plant Production 112: 63–67.

Zhang, J., R. Chiodini, A. Badr, and G. Zhang. 2011. The impact of next generation sequencing on genomics. J. Genet. Genomics 38: 95–109.