International Journal of Agricultural Technology 2023 Vol. 19(1): 257-276
Available online http://www.ijat-aatsea.com
ISSN 2630-0192 (Online)

Genetic diversity of arabica coffee varieties from the Royal Project Foundation using genotyping-by-sequencing (GBS)

Techanan, A.[1], Rumjuankiat, K.[1], Boonmee, K.[1], Chumpun, C.[2], Panchan, I.[2], Urairong, H.[1*] and Angkasith, P.[1,2]

[1]College of Agricultural Innovation and Food Technology, Rangsit University, Pathum Thani 12000, Thailand; [2]Royal Project Foundation, Chiang Mai 50200, Thailand.

Techanan, A., Rumjuankiat, K., Boonmee, K., Chumpun, C., Panchan, I., Urairong, H. and Angkasith, P. (2023). Genetic Diversity of Arabica Coffee Varieties from The Royal Project Foundation using Genotyping-by-Sequencing (GBS). International Journal of Agricultural Technology 19(1): 257-276.

Abstract Arabica coffee is one of the essential crops in the highland areas in Northern Thailand. Agro-morphological characterization, leaf rust resistance, and selected elite lines of these crops have been studied. However, little is known about the genetic diversity of Arabica coffee conserved by the Royal Project Foundation. This research used the genotyping-by-sequencing (GBS) method to investigate the genetic diversity of 23 Arabica coffee varieties. The results revealed that sequencing generated 1,945,389 average high-quality read tags per sample, of which 95.83% were successfully aligned to the reference genome. The total number of SNPs and InDels detected in the whole genome were 1,180,245 and 115,112 positions, respectively. The SNP dataset was used to analyze the pairwise similarity among populations based on the p-distance method, revealing narrow genetic diversity values from 0.14 to 0.56. The phylogenetic tree was divided into three main groups with a common ancestor. Moreover, the selected polymorphism of SNPs and InDels were developed to be molecular markers for coffee plant genetics studies. The outcome can evolve molecular markers and applications for plant breeding.

Keywords: Arabica coffee, Genotyping-by-Sequencing (GBS), SNPs, InDels

Introduction

Coffee is one of the most popular beverages in the world. Consequently, coffee beans are an important commercial crop for global coffee production. In addition, coffee beans and leaves have bioactive compounds and some antioxidant activities (Lestari *et al*., 2022). Generally, coffee is organized in the *Coffea* genus of the Rubiaceae family (Tran *et al*., 2016), which comprises several species, including *Coffea arabica* L. (Arabica), *Coffea canephora* var. *robusta* (Robusta), *Coffea excelsa* (Excelsa) and *Coffea liberica* (Liberica) var.

[*] **Corresponding Author**: Urairong, H.; **Email**: fongptt@yahoo.com

liberica. However, Arabica is the dominant species and is considered the best in taste and aroma (Mondragon and Santos, 2021; Tran *et al*., 2018). Furthermore, Arabica is an allotetraploid (2n = 4x = 44) derived from natural hybridization between *Coffea canephora* and *Coffea eugenioides* (Merot-L'anthoene *et al*., 2019). Arabica is a significant revenue generator for Thailand, the third largest producer in Southeast Asia - after Vietnam and Indonesia (Cheong *et al*., 2013; Phrommarat, 2019).

In Thailand, the identification and characterization of coffee have been using morphological features, which is inadequate due to environmental influences on the phenotype (Krishnan *et al*., 2021; Nasanit and Satayawut, 2015). Molecular markers can partition the environment from genetic effects on the phenotype, thus providing a higher degree of precision on the genetic relationship of the different genotypes (Guo *et al*., 2019). Recent advances in next-generation sequencing technologies have allowed the plant genome to be sequenced efficiently and economically. These may be used to directly detect single nucleotide polymorphisms (SNPs) in the whole genome of plants (Yang *et al*., 2020). Genotyping-by-sequencing (GBS) is a method to identify genetic variants of samples from rapid genotypes and reduces genome complexity by using restriction enzymes to split the genome into fragments sequenced on short-read sequencing platforms cost-effective. It generates large numbers of mutations for use in genetic analyses and genotyping with low cost, less sample handling, fewer PCR and purification stages, no reference sequence restrictions, and effective barcoding. Therefore, GBS is most commonly used to find a mutation of several complicated crops for developing plant breeding programs (He *et al*., 2014; Deschamps *et al*., 2012; Wang *et al*., 2020), such as oat (Huang *et al*., 2014), soybean (Wickland *et al*., 2017), maize (Wang *et al*., 2020), and coffee (Anagbogu *et al*., 2019). Consequently, next-generation sequencing technologies have been used to adapt molecular markers for coffee genotyping. (Alkimim *et al*., 2018; Marques *et al*., 2022).

In this study, we used genotyping-by-sequencing to analyze 23 coffee genotypes planted at the Royal Project Foundation in different regions of Northern Thailand, the primary coffee-production area in the country. The aims were to determine the genetic diversity among the Arabica coffee cultivars of the Royal Project Foundation, to validate SNPs and InDels detected from GBS in Arabica coffee, and to develop primers for detecting SNPs and InDels for investigating cultivars.

Material and methods

Plant materials

The leaf samples of twenty-four Arabica coffee cultivars were collected from the preservation field of in-situ conservation coffee plant genetic resources at the Mae Lod Royal Agricultural Research Station under the Royal Project Foundation. Cultivar names and sources of seed collection showed in Table 1. The leaf samples were preserved and stored at -80 ℃ after harvesting.

DNA Extraction

DNA of only young leaves was extracted by a modified CTAB method, according to Gargouri and Kacem (2018) .The quality of the DNA was assessed by 1% agarose gel electrophoresis. The concentration was then measured at 260/280 nm absorbance using a Nanodrop 1000 spectrophotometer (Thermo Fisher Scientific Inc., USA). Twenty-three diverse Arabica coffee cultivars, except Chiang Mai 80, were used for genotyping-by-sequencing. Sixteen Arabica coffee local cultivars were used to validate the SNPs and InDels region.

GBS library preparation and sequencing

The genomic DNA was sequenced by Novogene Co., Ltd. (Beijing, China) for genotyping-by-sequencing. DNA samples were digested with proper restriction enzymes *Msel* based on the *in silico* evaluation results to obtain a suitable marker density. The fragments were ligated with two barcoded adapters, either sticky end compatible with the primary digestion enzyme and the Illumina P5 or P7 universal sequence. Followed by PCR amplification, all the samples were pooled and size-selected for the required fragments to complete the library construction. The prepared DNA libraries were first checked using the Qubit 2.0 fluorometer to determine library concentration. After dilution to 1 ng/μl, the Agilent 2100 bioanalyzer was used to assess the insert size. Finally, the effective concentration of each library was estimated by the quantitative real-time PCR (qPCR). The qualified DNA libraries were pooled according to their effective concentration. Pair-end sequencing was performed on the Illumina platform, with a read length of 144 bp at each end. The sequenced raw data contains adapter contamination and low-quality reads which may increase the complexity of downstream analysis. Thus, the adapter sequence and low-quality reads were removed in the control step.

Table 1. List of cultivars/varieties and collection sources from the Royal Project Foundation in Northern Thailand

No.	Cultivars/Varieties	Sources
1	Catimor (H528/46)	Mae Lod Royal Agricultural Research Station, Mae Tang, Chiang Mai (19 °06'06.3"N 98 °46'23.6"E)
2	Mundo NoVo (H420/9)	
3	H373/24	
4	Caturra hybrid (H496/52)	
5	Congusta (Robusta x Congensis)	
6	Catimor (H.W. 26/7)	
7	Dwarf San Ramon	
8	Catimor	
9	Chiang Mai 80	
10	TT10-1	Teen Tok Royal Project Development Center, Mae On, Chiang Mai. (18 °52'00.8"N 99 °19'20.8"E)
11	TT8-5	
12	TT6-4	
13	TT5-1	
14	PMN2-8	Inthanon Royal Agricultural Station, Jom Thong, Chiang Mai (18 °32'33.5"N 98 °31'09.8"E)
15	PMN1-1	
16	MYN3-9	
17	ML2-6	Mae Lod Royal Agricultural Research Station, Mae Tang, Chiang Mai (19 °06'06.3"N 98 °46'23.6"E)
18	ML5-15	
19	PM1-7	Pa Miang Royal Project Development Center, Doi Saked, Chiang Mai (18 °59'03.5"N 99 °20'02.7"E)
20	PH1-5	
21	HH2-10	Mae La Noi Royal Project Development Center, Mae La Noi, Mae Hong Sorn (18 °20'29.1"N 98 °04'40.6"E)
22	HH1-13	
23	HH3-5	
24	HNK1-5	Huay Nam Khun Royal Project Development Center, Mae Sa Rauy, Chiang Rai (19 °34'31.4"N 99 °29'25.5"E)

Reference genome mapping, SNPs and InDels calling

For GBS, reads of each sample were sequenced and compared to the reference genome. The clean reads were aligned with the reference sequence from NCBI database (Reference genome, https://ftp.ncbi.nlm.nih.gov/genomes /all/annotation_releases/13443/100/GCF_003713225.1_Cara_1.0/GCF_003713 225.1_Cara_1.0genomic.fna.gz) through BWA (Burrows-Wheeler Aligner)

software (Li and Durbin, 2009). The mapping rate and coverage were counted according to the alignment results. The BAM files were handled by the SAMtools software package (Li *et al*., 2009), which was used to detect mutations. The individual SNP and InDel variations were detected using SAMtools with the following parameter: 'mpileup -m 2 -F 0.002 -d 1000'. These SNPs and InDels were annotated with the functional variation by ANNOVAR (Wang *et al*., 2010).

Phylogenetic analysis

The genetic distances were computed using the variation of the SNPs dataset of the whole genome derived from the GBS method using the p-distance method (Nei and Kumar, 2000). Then, the evolutionary history was inferred by the UPGMA method with a 1000 bootstrap test (Felsenstein, 1985). All evolutionary analyses were conducted in MEGA X (Kumar *et al*., 2018).

Selection of SNP positions and primer design

The position of the SNPs was chosen to be related to the population's polymorphism based on the GBS results and located in exon regions following the annotation. The primer for amplification and sequencing was designed by PyroMark Assay Design SW 2.0. The forward or reverse primer must be labeled with biotin as biotinylated PCR products will bind to the Streptavidin-Sepharose beads. We designed a PCR primer to cover each region where there are many mutations in one area. Thus, the sequencing primer may detect more than one mutation.

SNPs validation by pyrosequencing

The PCR amplification was set up and performed on a T100 Thermal cycler (Bio-Rad, USA) for the pyrosequencing analysis. The PCR mixture was prepared with 12.5 µl of 2x PyroMark PCR master mix, 2.5 µl of 10x CoralLoad concentrate, 1 µl of forward and reverse primers at a concentration of 5 µM, 0.8 µl of 40 ng/µl DNA template, and RNase-free water was added to a final reaction volume of 25 µl. PCR conditions were 95 ℃ for 15 min, followed by 45 cycles. The denaturation step was at 94 ℃ for 30 s, with primer annealing at 60 ℃ for 30 s and primer extension at 72 ℃ for 30 s. The final extension was 72 ℃ for 10 min. The PCR products were employed as the template being the pyrosequencing assay for SNP genotyping was carried out on a PyroMark Q48 Autoprep (QIAGEN GmbH, Germany).

Selection of InDel regions and primer design

Two requirements guided the selection of the InDel regions: (1) The population exhibits polymorphism from GBS results, and (2) The length of the insertion or deletion should be greater than eight bp to be detected by agarose gel electrophoresis. The primers for amplifying the InDel regions were designed under two restrictions: (1) The primers must be complementary to conserved target sequences and not be mismatched to other loci, and (2) The PCR products need to be between 100 and 300 bp for the polymorphism of the genotypes to be visible through the naked eye on a 3% MetaPhor agarose gel. The Primer3-Plus website was used to design primer sequences for the InDel regions.

InDels validation

A total volume of 20 µl of PCR reaction mix was prepared by 14 µl Dream Taq Green PCR master mix (Thermo Fisher Scientific, Inc., USA), 1 µl of forward and reverse primers at a concentration of 10 µM, 1.5 µl of 40ng/µl DNA template, and 2.5 µl of RNase-free water. The reaction mixture was initially denatured at 95 ℃ for 3 min; followed by ten cycles of amplification at 95 ℃ for 30 s, 60 ℃ for 30 s, and 72 ℃ for 40 s, and finally, 30 cycles of amplification at 95 ℃ for 30 s, the annealing temperature of roughly 53-55 ℃ for 30 s, depending on each pair of primers, and 72 ℃ for 40 s. Then the final extension was 72 ℃ for 7 min. The genomic DNA was amplified by a T100 Thermal cycler (Bio-Rad, USA). The PCR products were separated using 3% MetaPhor agarose gel (BMA, USA).

Results

Sequencing data

A collection of 23 *Coffea arabica* accessions was sequenced successfully by the Illumina HiSeq platform. It provided 7.018 G of raw data, of which 7.015 G (99.9%) was generated after filtering low-quality data. The clean data production for each sample ranged from 195,488,000 bp to 389,285,000 bp, indicating sufficient data production. The average of clean data is 305,016,173 bp per individual. The mapping rate of each sample ranges from 92.35% to 97.27%, and the Illumina-GBS pipeline identified 10,288,802 unique sequence tags. This result is in the qualified normal range and may serve in the subsequent variation detection and related analyses.

SNPs and InDels calling

The total SNPs and InDels variants detected from 23 cultivars of Arabica coffee were 1,180,245 and 115,112 positions, respectively. There are 69,422 positions of exonic SNPs; the mutation located in the exonic region or non-synonymous single nucleotide mutation can change the amino acid when occurred. In contrast, the mutations found in the intronic area were 109,686 positions. However, all the detected SNP positions can be divided into transition (809,711) and transversion (370,534) shown in Table 2. The DNA substitution mutations detected were separated into two types as follows: (1) transition was a point mutation that changes a purine nucleotide to another purine (A ↔ G) or a pyrimidine nucleotide to another pyrimidine (C↔ T), and (2) transversion was the substitution of a purine (Two rings) for a pyrimidine (One ring). When the T>C mutation appears on either double-strand, the A>G mutation will be found in the same position as the other chain. Therefore, the T>C and A>G mutations were classified into one category; for example, the T:A>C:G mutations include mutations from T to C and A to G. Other mutations were grouped similarly. In this study, SNPs mutation can be divided into six groups. The transition type of mutation was separated into two groups: T:A>C:G and C:G>T:A, which were the majority of SNPs found in 23 Arabica coffee genotypes. In addition, the transversion substitution mutation type was split into four groups: T:A>G:C, T:A>A:T, C:G>G:C, and C:G>A:T, which amount of it occurred lower than the transition type of SNPs. However, the Congusta had many SNPs mutations 2-3 times higher than the others in the study population (Figure 1). Moreover, the total insertion and deletion mutations (InDels) of nucleotides in 23 cultivars of genomic DNA were 74,147 and 40,965 positions, respectively. Meanwhile, the InDel positions in exonic and intronic were 2,039 and 14,258, respectively (Table 2). Additionally, the range length of InDel mutations was 1-21 base pairs (bp). One bp InDels were the majority mutation representing nearly 60% of the total, followed by three bp of InDels, which almost reached 20%, while another length of InDels 6, 9, 12, and 15 bp was under 10%. Interestingly, we found that the InDels nucleotide 1, 3, 6, 9, and 12 bp in the Arabica coffee genome occurred more than other sizes of InDels (Figure 2).

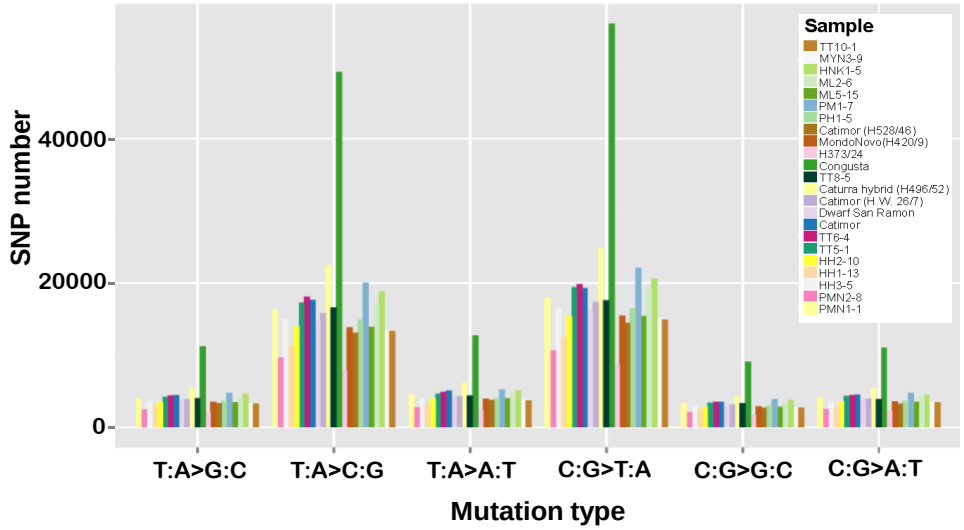

Figure 1. SNP mutation types distribution. The horizontal axis shows the type of mutation, whereas the vertical axis shows SNP numbers. Colors represent the different samples

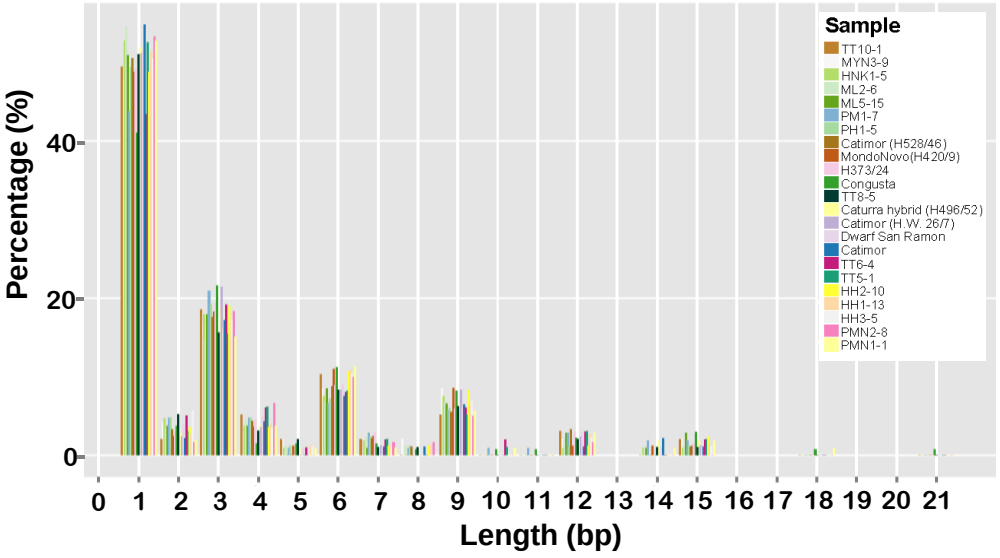

Figure 2. Length distribution of InDels. The horizontal axis shows the length (bp) of InDel, and the vertical axis shows the percentage (%). Colors represent the different samples

Table 2. Total of statistical SNPs and InDels detection with annotation

Category	SNP numbers	InDel numbers
Upstream	68,306	8,950
Exonic	69,422	2,039
Intronic	109,686	14,258
Splicing	242	102
Downstream	56,019	6,941
Upstream/Downstream	7,166	1,268
Intergenic	785,913	67,305
Other	83,491	14,249
Total	1,180,245	115,112
Transitions	809,711	-
Transversions	370,534	-
Insertion	-	74,147
Deletion	-	40,965

Upstream: Mutations within 1 kb upstream (Away from transcription start site) of the gene.
Exonic: Mutations located in the exonic region
Intronic: Mutations located in the intronic region.
Splicing: SNPs located in the splicing site (2 bp range of the intron/exon boundary).
Downstream: Mutations within 1 kb downstream (Away from transcription termination site) of the gene region.
Upstream/Downstream: Mutations located within the < 2 kb intergenic region are in 1 kb downstream or upstream of the genes.
Intergenic: Mutations located within the > 2 kb intergenic region.

Phylogenetic analysis

Genetic distance and phylogenetic tree were used to estimate the genetic relationships between or within the population. The distance matrix was computed using the p-distance method based on the variation of single nucleotide substitutions among 23 genotypes of the Arabica coffee. All ambiguous positions were removed for each sequence pair (Pairwise deletion option). There was a total of 66219 positions in the genetic similarity data set, which was used to reconstruct the phylogenetic tree by the UPGMA method. The result of the p-distance matrix revealed values ranging from 0.14 to 0.56. Nevertheless, the genetic dissimilarity of the Congusta shows it was divergent by a p-distance value of around 0.53 to 0.56 compared to all study genotypes (Figures 3 and 4B), so this divergent cultivar was not included in the reconstruction of the phylogenetic tree. The genotypes were grouped into three main clusters I, II, and III. Cluster I comprise two sub-clusters. Sub-cluster I-1

consists of Dwarf San Ramon, PMN2-8, MYN3-9, Catimor (H528/46), H373/24, HH2-10, and TT5-1. We observed that closely related genotypes were the pair of PMN2-8 against Dwart San Ramon and PMN2-8 against MYN3-9 by a distance value of 0.14 (Figures 3 and 4A). Sub-cluster I-2 included TT10-1, HH1-13, Mundo Novo (H420/9), and PMN1-1. Meanwhile, Cluster II contained three sub-clusters. Catimor (H.W. 26/7) belongs to sub-cluster II-1. Sub-cluster II-2 comprises ML2-6, ML5-15, and Catimor. Sub-cluster II-3 constitutes PH1-5, Caturra hybrid (H496/52), PM1-7, HNK1-5, TT8-5, and HH3-5. Besides, the TT6-4 was separated from other cultivars and grouped as Cluster III (Figure 4A).

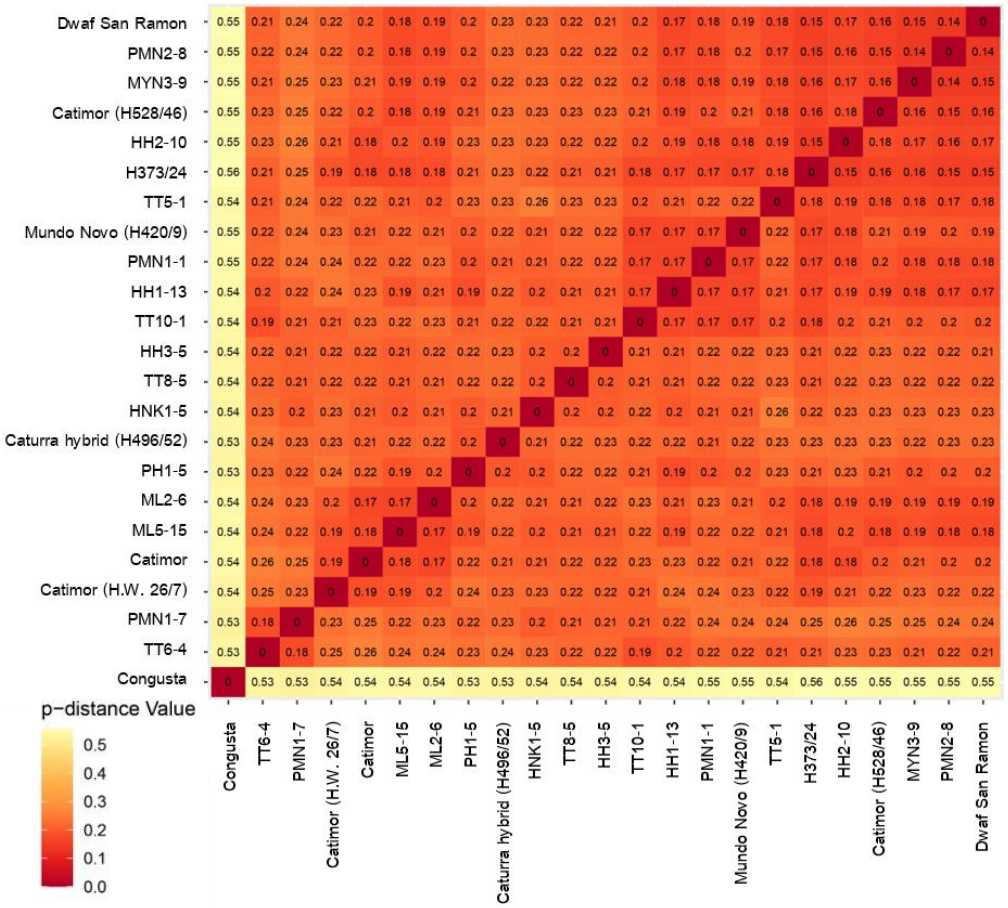

Figure 3. Estimates of evolutionary genetic distance between sequences of 23 cultivated Arabica coffee using the p-distance method. The p-distance value is the proportion (p) of nucleotide sites as which the two compared sequences differ

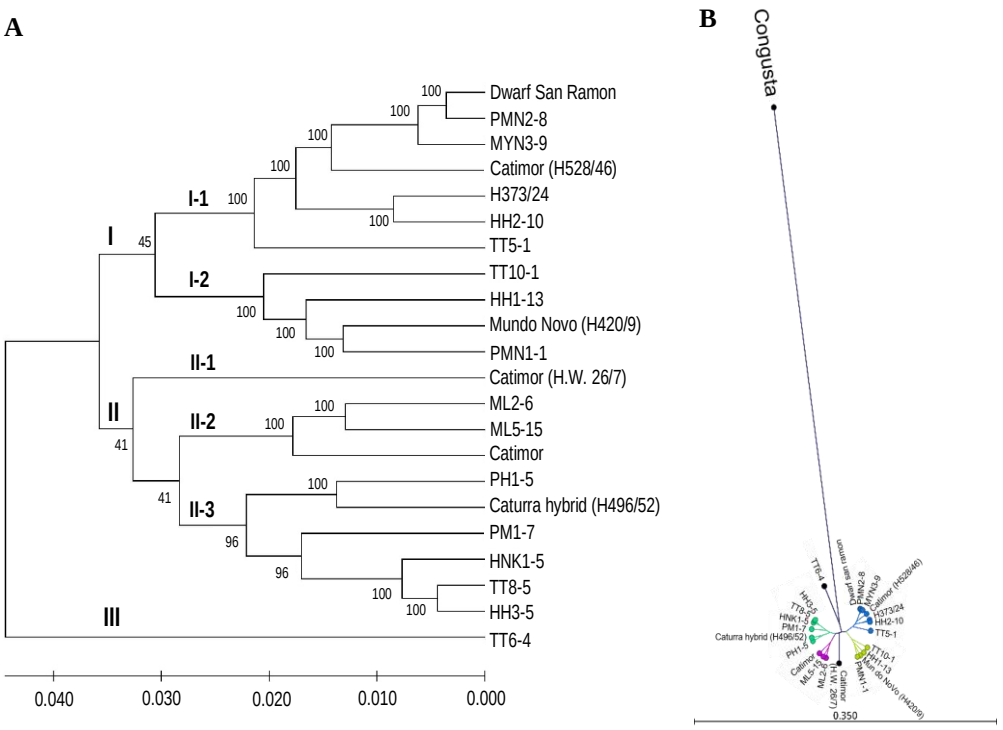

Figure 4. The phylogenetic tree of cultivated Arabica coffee from the Royal Project Foundation is based on the p-distance method obtained from nucleotide sequences using the UPGMA method. The bootstrap test used 1000 replications. The support of the nodes is indicated by numbers displayed on the terminal branches: (A) The phylogenetic tree of the 22 cultivars was constructed without an outgroup. (B) The phylogenetic tree of the 23 cultivars was constructed with an outgroup (Congusta variety)

Validation of SNP positions using pyrosequencing

Using the SNPs from GBS analysis as a marker for genotyping Arabica coffee, the exonic SNPs with polymorphism in a studied sample were chosen for verification by pyrosequencing assays. There are three different genes as follows: (1) elongation factor-like GTPase 1 pseudogene (NC_039899.1), (2) probably inactive leucine-rich repeat receptor-like protein kinase At5g48380 (NC_039903.1), and (3) dynamin-related protein 4C-like (NC_039915.1). The nucleotide sequence on chromosome numbers NC_039899.1, NC_039903.1, and NC_039915.1 comprises four, nine, and seven SNP positions, respectively, according to selected criteria. The newly developed pyrosequencing primers were initially applied to validate the earlier SNPs (Table 3) in the 16 Arabica

coffee genotypes. Validating among twenty SNP positions with pyrosequencing assay shows eight polymorphism positions at columns 1, 2, 5, 6, 8, 9, 16, and 17, as the same report by GBS. The rest was monomorphism. We found that the SNP position in column no.16 can be used as a marker to differentiate the TT8-5 genotype from another group. Furthermore, a set of SNP positions at columns no.6, 8, 9, and 17 were identified in the PMN2-8. The HH3-5 genotype was classified by SNP positions in columns 9 and 17, whereas columns 5 and 17 were used to separate the ML5-15 from other genotypes (Table 4). However, more primers were needed to differentiate each cultivar in this study effectively.

Table 3. NCBI RefSeq, SNP positions, PCR primer, and sequencing primer for pyrosequencing assays

NCBI RefSeq	SNP Positions	PCR primer	Sequencing primer
NC_039899.1	36192541	F: 5'-TATGTTAACGAAGC TGCTTGATG-3' R: 5'-Biotin-TGTCCCCTAA TTAAAACAGGAGAA-3'	5'-CGAGAGGTAGC ATTGTG-3'
	36192546		
	36192622		5'-ATGGAGATTCA GAGGC-3'
	36192646		
NC_039903.1	10315546	F: 5'-CGGCCATTAATTCA AGGCAT-3'	5'-AAGGCATCTAG CCCT-3'
	10315637	R: 5'-Biotin-TTTAGTCCCA TCCGCGTAGG-3'	5'-GGAGTCCTTGG ATGAT-3'
	10315648		
	10315651		
	10315672		5'-AGCTGGAATTA TGACAAC-3'
	10315680		
	10315715		5'-GGGTGTTGAAT GCTG-3'
	10315722		
	10315740		
NC_039915.1	2704066	F: 5'-AGCCCGTCTGCATA TACCGTAAA-3' R: 5'-Biotin-CATGGAAAC CGAAATGGTGAAT-3'	5'-AAAAACAACAA AGCAATAGT-3'
	2704069		
	2704081		
	2704099		
	2704162		5'-CTCTGCAACAAT CTCTTTA-3'
	2704184		
	2704215		5'-CTGTAACGTTTA CCAG-3'

F: Forward primer, R: Reverse primer

Table 4. Single nucleotide polymorphisms (SNPs) detection and allele ratios of heterozygous mutation by pyrosequencing among three genes from sixteen Arabica coffee cultivars

NCBI RefSeq	NC_039899.1				NC_039903.1									NC_039915.1						
Gene	Elongation factor-like GTPase 1 pseudogene				Probably inactive leucine-rich repeat receptor-like protein kinase At5g48380									Dynamin-related protein 4C-like						
Position	36192541 1/	36192546 1/	36192622	36192646	10315546 1/	10315637 1/	10315648	10315651 1/	10315672 1/	10315680	10315715	10315722	10315740	2704066	2704069	2704081 1/	2704099 1/	2704162	2704184	2704215
Column No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
SNP	T/A	G/A	A/C	T/C	C/T	T/C	A/G	A/T	A/C	G/A	C/A	A/G	C/T	G/A	T/C	G/C	T/G	C/G	C/G	G/A
TT10-1	T/A (54:46)	G/A (53:47)	A/C (56:44)	T/C (51:49)	C/T (52:48)	T/C (59:41)	A	A/T (64:36)	A	G	C	A	C	G	T	G	T/G (74:26)	C	C	G
TT8-5	A	A	A/C (51:49)	T/C (50:50)	C	T	A	A	A	G	C	A	C	G	T	G/C (48:52)	T/G (40:60)	C	C	G
TT6-4	A	A	A/C (37:63)	T/C (36:64)	C	T	A	A	A	G	C	A	C	G	T	G	T/G (70:30)	C	C	G
TT5-1	T/A (52:48)	G/A (50:50)	A/C (52:48)	T/C (49:51)	C	T	A	A	A	G	C	A	C	G	T	G	T/G (73:27)	C	C	G
HH2-10	T/A (53:47)	G/A (57:43)	A/C (55:45)	T/C (52:48)	C	T	A	A	A	G	C	A	C	G	T	G	T/G (74:26)	C	C	G
HH1-13	T/A (51:49)	G/A (52:48)	A/C (58:42)	T/C (50:50)	C	T	A	A	A	G	C	A	C	G	T	G	T/G (62:38)	C	C	G
HH3-5	T/A (49:51)	G/A (49:51)	A/C (51:49)	T/C (51:49)	C	T	A	A	A/C (73:27)	G	C	A	C	G	T	G	G	C	C	G
PMN2-8	T/A (49:51)	G/A (56:44)	A/C (55:45)	T/C (50:50)	C	T	A	A/T (73:27)	A/C (67:33)	G	C	A	C	G	T	G	T	C	C	G
PMN1-1	T/A (51:49)	G/A (48:52)	A/C (55:45)	T/C (50:50)	C	T	A	A	A	G	C	A	C	G	T	G	T/G (57:43)	C	C	G
MYN3-9	T/A (49:51)	G/A (53:47)	A/C (53:47)	T/C (51:49)	C	T	A	A	A	G	C	A	C	G	T	G	G	C	C	G
HNK1-5	T/A (52:48)	G/A (50:50)	A/C (53:47)	T/C (51:49)	C	T	A	A	A	G	C	A	C	G	T	G	T/G (63:37)	C	C	G
ML2-6	T/A (51:49)	G/A (51:49)	A/C (53:47)	T/C (50:50)	C	T	A	A	A	G	C	A	C	G	T	G	T/G (70:30)	C	C	G
ML5-15	T/A (52:48)	G/A (48:52)	A/C (52:48)	T/C (50:50)	C/T (74:26)	T	A	A	A	G	C	A	C	G	T	G	G	C	C	G
PM1-7	A	A	A/C (53:47)	T/C (47:53)	C	T	A	A	A	G	C	A	C	G	T	G	T/G (74:26)	C	C	G
PH1-5	T/A (50:50)	G/A (55:45)	A/C (52:48)	T/C (52:48)	C	T	A	A	A	G	C	A	C	G	T	G	T	C	C	G
Chiang Mai 80	T/A (55:45)	G/A (59:41)	A/C (50:50)	T/C (47:53)	C/T (72:28)	T/C (66:34)	A	A/T (73:27)	A	G	C	A	C	G	T	G	T/G (71:29)	C	C	G

1/: Polymorphic position

Validation of InDel positions

From GBS analysis, the polymorphic InDels of these samples located in the exonic were chosen as a marker for genotyping Arabica coffee using PCR and gel electrophoresis. There are seven different regions as follows: (1) alcohol dehydrogenase-like 1, (2) cytochrome P450 94A1-like, (3) uncharacterized LOC113726705, (4) CRIB domain-containing protein RIC6-like, (5) la-related protein 1A, (6) protein RER1A-like, and (7) splicing factor 3B subunit 1 pseudogene. The above InDels region was used to develop a set of markers to differentiate 16 Arabica coffee genotypes by designing a primer for each region as indicated in Table 5. The collection of primers was categorized based on the PCR amplicons from agarose gel electrophoresis. Of these primers, four pairs were monomorphic markers, and three were polymorphic markers that can identify differences in the 16 coffee genotypes. The InDCA4 primer amplified a single identical allele of TT8-5, whereas the other genotypes were more than one allele. Moreover, the TT6-4 and PMN1-7 were identified to get out of other genotypes, with the InDCA5 primer showing different alleles. In contrast, The InDCA7 primer could separate the five genotypes from sample groups, including TT10-1, TT8-5, TT6-4, TT5-1, and PMN1-7 (Figure 5).

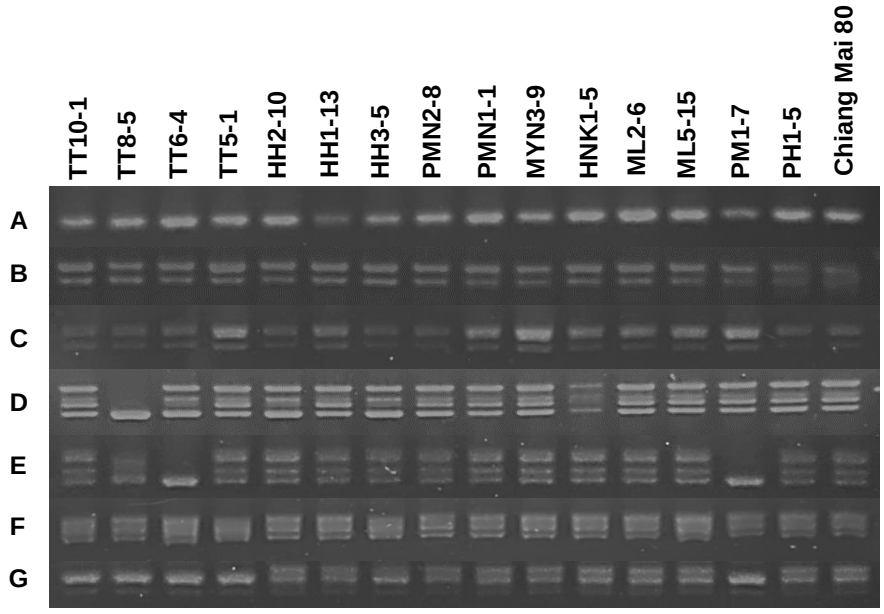

Figure 5. Genotypes of the 16 cultivars at seven InDel markers as follows: InDCA1 (A), InDCA2 (B), InDCA3 (C), InDCA4 (D), InDCA5 (E), InDCA6 (F), and InDCA7 (G)

Table 5. Primer sequences for validation of the InDel regions

Primer name	Forward primer (5'-3')	Reverse primer (5'-3')
InDCA1	GTGAGGCACCCAAGGTAGAA	GGAAAGCCATTGCAGCATAA
InDCA2	CCTTTCCTCCTAAAATTAC	CGTTTAGCCTCCCAAATCAA
InDCA3	GCTGTTTGAGTTTGTGGGCTC	GTTCTTAAACCAACAAACCCG
InDCA4	GCACCAATTCTGTTCCACTTC	GCTTTCTTGAGTGCCTGGAT
InDCA5	CGACGTATCTAGTTGATTCAGC	TTCCAGACCGTACAGACTGC
InDCA6	AGGAGGGAGGCAGTAAAGGA	GCAGATGATACTTTTAGTCGG
InDCA7	GTAGCTGGTTCTATTATTCTGG	CCCTGATCGAAACCCTAACC

Discussion

Genotyping-by-sequencing (GBS) has been used as an effective tool for the determination of the genetic diversity in many crop plants, including melon (Pavan *et al*., 2017), pepper (Pereira-Dias *et al*., 2019), coffee (Anagbogu *et al*., 2019), and maize (Wang *et al*., 2020), etc. This approach has been advantageous because it provides simultaneous analysis of the whole genome of all samples to discover polymorphic SNP positions highly relevant to the genotype (Wang *et al*., 2020).

The SNPs is used for the positions of the 23 Arabica coffee nucleotides derived from the GBS method to reconstruct the phylogenetic tree, which included Congusta, the progeny of interspecific hybridization between *C. canephora* and *C. congensis* (van Boxtel and Berthouly, 1996) as an outgroup in comparison with ingroup population regarding genetic distance or degree of genetic difference. The result showed clear visualization that the 22 ingroup cultivars had very narrow genetic diversity indicated by the p-distance value of 0.14-0.26; meanwhile, the high p-distance value (0.53-0.56) and the phylogenetic tree pattern confirmed that Congusta was suitable to be used as an outgroup. This phylogenetic tree comprised 3 clusters indicating that the member in each cluster shares the most recent common ancestor. In detail, cluster I included two sub-clusters. Sub-cluster I-1, p-distance value showed PMN2-8 more similar to Dwarf San Ramon, and HH2-10 more similar to H373/24 than another member in this sub-group. In sub-cluster I-2, PMN1-1 and Mundo Novo (H420/9) were more related than HH1-13 and TT5-1. Considering that the most recent common ancestor of all member in cluster I may be similar to the progenitor of some member have been reported as follow, Dwarf San Ramon originated from a dwarf mutant of Bourbon, Catimor (H528/46) derived from the cross between Catuai Amarelo and H.W. 26/13, H373/24 is a hybrid cross between Bourbon 43-7xRP 13 (10/1) and H.W. 26/9, and Mundo Novo is a hybrid in natural between Red Bourbon and Typica

(Etienne *et al*., 2002; Gichuru, 2007; Maltsbarger, 2011). Within cluster II, there were ten coffee cultivars divided into three sub-cluster. The famous cultivar and their genetic background were as follows, Catimor was a cultivar originating from a cross between Hibrido de Timor and Caturra (Zhang *et al*., 2012), Catimor (H.W. 26/7) derived from the hybridization between Hibrido de Timor 832/1 and Caturra Vermelho (Silva *et al*., 2022), and Caturra hybrid (H496/52) was the offspring of the cross between Caturra Amarello and H51/5 (genetic background from Kent and Agaro). In conclusion, the member cultivars within cluster I evolved from the main common ancestor of Bourbon, Mundo Novo, Catuai, and Typica. In Cluster II, all member cultivars have a genetic background from Hibrido de Timor, Caturra, and others. Concerning this approach, many crop plants have been reported as follows, Pereira-Dias *et al*. (2019) also classified pepper by creating a phylogenetic tree using the GBS approach. While Wickland *et al*. (2017) reported that this method had missing data due to analyze many short gun sequences. However, it can provide SNPs and InDels regions for utilization in classifying plant cultivars (Deschamps *et al*., 2012), moreover, if SNP was located in essential genes such as fragrance genes and other dominant genes. It can be used to select cultivars with unique properties for development in plant breeding programs (He *et al*., 2014).

Simultaneously, the number of SNPs and InDels obtained from the GBS method can be streamlined in studying the genetic relationship of several plants (Li *et al*., 2016; Shen *et al*., 2017; Roy and Lachagari, 2017). It is popular to apply SNPs in gene mutation research to improve the essential properties of plants (Kim *et al*., 2016; Hu *et al*., 2020; Tran *et al*., 2018). The study of Zhu *et al*. (2015) used SNPs and InDels regions to determine the genetic relationships of *Poncirus*. They successfully used variant SNPs and InDels to distinguish the population with 11 mutation positions. In the current study, GBS method is applied to determine the genetic diversity of 23 Arabica coffee cultivars collected from The Royal Project Foundation. The result agreed this method has high-throughput efficiency, relatively low cost, and time-saving to discover SNP and InDel positions among 23 coffee genomes. Furthermore, 20 SNP positions were verified using the pyrosequencing assay. The results showed that SNPs could differentiate TT8-5 genotypes from other cultivars in the population. Although, the evolutionary relationship in a phylogenetic tree showed that the TT8-5 genotype was closely related to the HH3-5 genotype. Moreover, the validation of 7 InDel positions showed that the InDCA4 primer can differentiate TT8-5 genotypes from this group. Thus, the results were shown that SNPs and InDels could be utilized to determine coffee genotypes.

So, the phylogenetic tree is not only provided the genetic similarities or differences among varieties or species but also gave scientists gain a better understanding of how varieties or species evolved from a common ancestor.

Acknowledgments

We express our gratitude to the Royal Project Foundation and College of Agricultural Innovation and Food Technology, Rangsit University, Thailand, for sample collection and funding support.

References

Anagbogu, C. F., Bhattacharjee, R., Ilori, C., Tongyoo, P., Dada, K. E., Muyiwa, A. A., Gepts, P. and Beckles, D. M. (2019). Genetic diversity and re-classification of coffee (*Coffea canephora* Pierre ex A. Froehner) from South Western Nigeria through genotyping-by-sequencing-single nucleotide polymorphism analysis. Genetic Resources and Crop Evolution, 66:685-696.

Alkimim, E. R., Caixeta, E. T., Sousa, T. V., da Silva, F. L., Sakiyama, N. S. and Zambolim, L. (2018). High-throughput targeted genotyping using next-generation sequencing applied in *Coffea canephora* breeding. Euphytica, 214:1-18.

Cheong, M. W., Tong, K. H., Ong, J. J. M., Liu, S. Q., Curran, P. and Yu, B. (2013). Volatile composition and antioxidant capacity of Arabica coffee. Food Research International, 51:388-396.

Deschamps, S., Llaca, V. and May, G. D. (2012). Genotyping-by-sequencing in plants. Biology (Basel), 1:460-483.

Etienne, H., Anthony, F., Dussert, S., Fernandez, D., Lashermes, P. and Bertrand, B. (2002). Biotechnological applications for the improvement of coffee (*Coffea arabica* L.). In Vitro Cellular & Developmental Biology - Plant, 38:129-138.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. Evolution, 39:783-791.

Gargouri, H. and Kacem, H. H. (2018). Evaluation of alternative DNA extraction protocols for the species determination in turkey salami authentication tests. International Journal of Food Properties, 21:733-745.

Gichuru, E. K. (2007). Characterization of genetic resistance to Coffee Berry Disease (*Colletotrichum kahawae* Waller and Bridge) in Arabica coffee (*Coffea arabica* L.) that is introgressed from *Coffea canephora* Pierre. (Doctoral dissertation). University of Nairobi, Nairobi.

Guo, G., Zhang, G., Pan, B., Diao, W., Liu, J., Ge, W., Gao, C., Zhang, Y., Jiang, C. and Wang, S. (2019). Development and application of InDel markers for *Capsicum* spp. based on whole-genome re-sequencing. Scientific Reports, 9:1-14.

He, J., Zhao, X., Laroche, A., Lu, Z.-X., Liu, H. and Li, Z. (2014). Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Frontiers in Plant Science, 5:1-8.

Huang, Y.-F., Poland, J. A., Wight, C. P., Jackson, E. W. and Tinker, N. A. (2014). Using genotyping-by-sequencing (GBS) for genomic discovery in cultivated oat. PloS one, 9:1-16.

Hu, W., Zhou, T., Wang, P., Wang, B., Song, J., Han, Z., Chen, L., Liu, K. and Xing, Y. (2020). Development of whole-genome agarose-resolvable LInDel markers in Rice. Rice, 13:1-11.

Kim, S. R., Ramos, J., Ashikari, M., Virk, P. S., Torres, E. A., Nissila, E., Hechanova, S. L., Mauleon, R. and Jena, K. K. (2016). Development and validation of allele-specific SNP/indel markers for eight yield-enhancing genes using whole-genome sequencing strategy to increase yield potential of rice, *Oryza sativa* L. Rice, 9:1-17.

Krishnan, S., Matsumoto, T., Nagai, C., Falconer, J., Shriner, S., Long, J., Medrano, J. F. and Vega, F. E. (2021). Vulnerability of coffee (*Coffea* spp.) genetic resources in the United States. Genetic Resources and Crop Evolution, 68:2691-2710.

Kumar, S., Stecher, G., Li, M., Knyaz, C. and Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. Molecular Biology and Evolution, 35:1547-1549.

Lestari, W., Hasballah, K., Listiawan, M. Y. and Sofia, S. (2022). Identification of antioxidant components of Gayo Arabica Coffee Cascara using the GC-MS method. 11th AIC 2021 on Earth and Environmental Science, Banda Aceh, Indonesia.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25:2078-2079.

Li, D., Zeng, R., Li, Y., Zhao, M., Chao, J., Li, Y., Wang, K., Zhu, L., Tian, W.-M. and Liang, C. (2016). Gene expression analysis and SNP/InDel discovery to investigate yield heterosis of two rubber tree F1 hybrids. Scientific Reports, 6:1-12.

Maltsbarger, R. D. (2011). A structural model of the international coffee sector: An econometric investigation. (Master Thesis). University of Missouri, Columbia.

Marques, I., Gouveia, D., Gaillard, J.-C., Martins, S., Semedo, M. C., Lidon, F. C., DaMatta, F. M., Ribeiro-Barros, A. I., Armengaud, J. and Ramalho, J. C. (2022). Next-generation proteomics reveals a greater antioxidative response to drought in *Coffea arabica* than in *Coffea canephora*. Agronomy, 12:1-20.

Merot-L'anthoene, V., Tournebize, R., Darracq, O., Rattina, V., Lepelley, M., Bellanger, L., Tranchant-Dubreuil, C., Coulee, M., Pegard, M., Metairon, S., Fournier, C., Stoffelen, P., Janssens, S. B., Kiwuka, C., Musoli, P., Sumirat, U., Legnate, H., Kambale, J. L., Ferreira da Costa Neto, J., Revel, C., de Kochko, A., Descombes, P., Crouzillat, D. and Poncet, V. (2019). Development and evaluation of a genome-wide coffee 8.5K SNP

array and its application for high-density genetic mapping and for investigating the origin of *Coffea arabica* L. Plant Biotechnology Journal, 17:1418-1430.

Mondragon, M. V. and Santos, D. M. C. (2021). Sequence analysis of select Terpene synthase genes in Philippine coffee. Philippine Journal of Science, 150:1549-1571.

Nasanit, R. and Satayawut, K. (2015). Microbiological study during coffee fermentation of *Coffea arabica* var. *Chiangmai 80* in Thailand. Kasetsart Journal - Natural Science, 49:32-41.

Nei, M. and Kumar, S. (2000). Molecular evolution and phylogenetics. Oxford University Press, New York.

Pavan, S., Marcotrigiano, A. R., Ciani, E., Mazzeo, R., Zonno, V., Ruggieri, V., Lotti, C. and Ricciardi, L. (2017). Genotyping-by-sequencing of a melon (*Cucumis melo* L.) germplasm collection from a secondary center of diversity highlights patterns of genetic variation and genomic features of different gene pools. BMC Genomics, 18:1-10.

Pereira-Dias, L., Vilanova, S., Fita, A., Prohens, J. and Rodr ́guez-Burruezo, A. (2019). Genetic diversity, population structure, and relationships in a collection of pepper (*Capsicum* spp.) landraces from the Spanish centre of diversity revealed by genotyping-by-sequencing (GBS). Horticulture Research, 6:1-13.

Phrommarat, B. (2019). Life cycle assessment of ground coffee and comparison of different brewing methods: A case study of organic Arabica coffee in Northern Thailand. The Environment and Natural Resources Journal, 17:96-108.

Roy, S. C. and Lachagari, V. B. R. (2017). Assessment of SNP and InDel variations among rice lines of Tulaipanji x Ranjit. Rice Science, 24:336-348.

Shen, C., Jin, X., Zhu, D. and Lin, Z. (2017). Uncovering SNP and indel variations of tetraploid cottons by SLAF-seq. BMC Genomics, 18:1-13.

Silva, M. D. C., Guerra-Guimar ães, L., Diniz, I., Loureiro, A., Azinheira, H., Pereira, A. P., Tavares, S., Batista, D. and V árzea, V. (2022). An overview of the mechanisms involved in coffee-*Hemileia vastatrix* interactions: Plant and pathogen perspectives. Agronomy, 12:1-30.

Tran, H. T. M., Lee, L. S., Furtado, A., Smyth, H. and Henry, R. J. (2016). Advances in genomics for the improvement of quality in coffee. Journal of the Science of Food and Agriculture, 96:3300-3312.

Tran, H. T. M., Ramaraj, T., Furtado, A., Lee, L. S. and Henry, R. J. (2018). Use of a draft genome of coffee (*Coffea arabica*) to identify SNPs associated with caffeine content. Plant Biotechnology Journal, 16:1756-1766.

van Boxtel, J. and Berthouly, M. (1996). High frequency somatic embryogenesis from coffee leaves. Plant Cell, Tissue and Organ Culture, 44:7-17.

Wang, K., Li, M. and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research, 38:1-7.

Wang, N., Yuan, Y., Wang, H., Yu, D., Liu, Y., Zhang, A., Gowda, M., Nair, S. K., Hao, Z., Lu, Y., San V. F., Prasanna, B. M., Li, X. and Zhang, X. (2020). Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding. Scientific Reports, 10:1-12.

Wickland, D. P., Battu, G., Hudson, K. A., Diers, B. W. and Hudson, M. E. (2017). A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. BMC Bioinformatics, 18:1-12.

Yang, T.-Y., Gao, T.-X., Meng, W. and Jiang, Y.-L. (2020). Genome-wide population structure and genetic diversity of Japanese whiting (*Sillago japonica*) inferred from genotyping-by-sequencing (GBS): Implications for fisheries management. Fisheries Research, 225:1-7.

Zhang, H., Li, J., Zhou, H., Chen, Z., Song, G., Peng, Z., Pereira, A. P., Silva, M. C. and Varzea, V. M. P. (2012). Arabica coffee production in the Yunnan province of China. 24th International Conference on Coffee Science (ASIC), San José, Costa Rica.

Zhu, S., Wang, F., Shen, W., Jiang, D., Hong, Q. and Zhao, X. (2015). Genetic diversity of *Poncirus* and phylogenetic relationships with its relatives revealed by SSR and SNP/InDel markers. Acta Physiologiae Plantarum, 37:1-11.

(Received: 10 September 2022, accepted: 30 December 2022)