Emirates Journal of Food and Agriculture. 2023. 35(3): 180-188
doi: 10.9755/ejfa.2023.v35.i3.3038
http://www.ejfa.me/

RESEARCH ARTICLE

Whole chloroplast genome sequences contribute to phylogenetic relatedness and cultivar identification in cacao (*Theobroma cacao* L.)

Nur Kholilatul Izzah[1], Reflinur[2]*, Hyun Seung Park[3], Jee Young Park[3], Tae-Jin Yang[3]

[1]Research Center for Horticultural and Estate Crops, Research Organization for Agriculture and Food, National Research and Innovation Agency, Cibinong Science Center, West Java, Indonesia, [2]Research Center for Genetic Engineering, Research Organization for Life Sciences and Environment, National Research and Innovation Agency, Cibinong Science Center, West Java, Indonesia, [3]Department of Agriculture, Forestry, and Bioresources, College of Agriculture and Life Sciences, and Plant Genomics and Breeding Institute, Seoul National University, Seoul, 151-921, Republic of Korea

ABSTRACT

Cacao (*Theobroma cacao* L.) is considered as economic importance crop playing a major role as source of chocolate industries for production of both chocolate candy and cocoa butter. Despite the advantages of cacao in industrial sector, understanding the global structure of cacao chloroplast genome plays a crucial role in explaining phylogenetic relationships and cultivar identification. This study aimed to perform phylogenetic analysis of cacao genotypes and develop DNA barcode markers using cacao chloroplast sequences. Chloroplast genome sequencing of two cacao genotypes (DR-1 and Sca-12) was conducted by the Illumina Miseq platform. Phylogenetic analysis of 12 cacao genotypes including two genotypes sequenced in this study (DR-1 and Sca-12) and ten genotypes previously sequenced (HQ336404, EET-64, ICS-01, ICS-06, ICS-39, Pentagonum, Sca-6, Stahel, Amelonado, and Criollo-22) showed a clear separation between bulk and fine types. This result demonstrated the usefulness of chloroplast sequences in revealing phylogenetic relatedness. Based on comparative chloroplast genome analysis of two cacao genotypes, DR-1 and Sca-12, three insertion/deletion (InDel) markers named as Theca_indel01, Theca_indel02, and Theca_indel03 which designed from the regions of *trnA-UGC*-rrn23, *trnK-UUU-rps16*, and *rps16* intron, respectively were successfully developed to reveal barcode system for cacao genetic discrimination. Overall, these findings would provide valuable insight into the chloroplast genome structure of cacao plant, as well as information about the benefits of chloroplast sequences for constructing phylogenetic relationships and developing DNA barcode markers.

Keywords: Cacao; Chloroplast genome; DNA barcode markers; Phylogenetic relationships

INTRODUCTION

Cacao (*Theobroma cacao* L.) is a perennial plant that commonly found in tropical countries with its center of origin in the Andes mountains, South America (Cheesman, 1944). Cultivated cacao has three varietal groups including Criollo, Forastero and Trinitario. Of which, these three cacao types showed different characteristics. Criollo type has the best quality of beans, whereas Forastero type usually more robust and resistant to some diseases. Another type is Trinitario, the hybrid of Criollo and Forastero, which combine the best traits from both parents (Kane et al., 2012). All these three varietal groups are important germplasm sources for cacao breeding program. Variations on cacao genotypes are needed by plant breeders to develop new high-yielding varieties. The advanced technologies in the field of DNA sequencing have enabled to understand variations within nuclear genome.

Chloroplast is active metabolic systems that contain numerous important proteins, especially those involved in the photosynthesis process (Daniell et al., 2016). Chloroplast genome in many embryophytes is characterized by a circular DNA molecules that ranges from 120 kb to 160 kb and contains three main structures, known as SSC (small single copy), LSC (large single copy), and IR (inverted repeats) (Chumley et al., 2006). The genome of chloroplast contains conserved genes useful for plant life processes as well as high variable regions that would be valuable in long-term research (Nock et al., 2011). The

***Corresponding author:**
Reflinur, Research Center for Genetic Engineering, Research Organization for Life Sciences and Environment, National Research and Innovation Agency, Cibinong Science Center, West Java, Indonesia. **E-mail:** reflinur@yahoo.com

Received: 09 September 2022; **Accepted:** 27 February 2023

region of chloroplast genomes that has high frequency of evolution is known to be important for phylogenetic relatedness and DNA barcoding, particularly for closely related species (Dong et al., 2012). The availability of many regions that have high evolutionary rates would be helpful to carry out such analysis. Fortunately, the complete chloroplast genomes in many plants species are now available since the successful sequencing of chloroplast in tobacco and liverwort (Ohyama et al., 1986; Shinozaki et al., 1986). Furthermore, advances in next-generation sequencing technology make the process of sequencing cacao chloroplast from total DNA more effective and efficient due to several advantages, such as high-throughput technology, save more time, and low expense (Liu et al., 2013).

Utilization of protein-coding genes sequences generated from *Artemisia* chloroplast genome has been applied in the phylogenetic tree construction for several species belong to Asteraceae family. In addition, these sequences were also used to develop seven InDel-based DNA barcodes that succeeded in identifying three species of *Artemisia*, including *Artemisia capillaris, A. gmelinii, and A. fukudo*, which are known to have a similar appearance (Lee et al. 2022). On the other hand, genomic studies have been developing rapidly in cacao. Formerly, two complete chloroplast (cp) DNA sequences obtained from SCA-6 genotype and an unknown genotype were publicly available (Jansen et al., 2011). Hence, such information would be beneficial to explore possible sequences in developing molecular marker applicable for both cultivar discrimination and phylogenetic analysis. Another study has applied ultra-barcoding (UBC) approach, a method that uses longer DNA-barcoding loci in ribosomal DNA and plastid to examine complete plastomes and nuclear rDNA sequences obtained from *T. cacao* genotypes as well as its related species (*T. grandiflorum*). Of which, all cacao genotypes examined were successfully differentiated. This result demonstrated feasibility and effectiveness of the UBC method to discriminate cacao varieties and even individual cacao genotypes (Kane et al., 2012). Recent studies have been successfully utilized DNA barcode markers designed from indels and transversion of the *trnH-psbA* as one of noncoding spacer present in the genome of cacao chloroplast which is able to differentiate cacao genotypes (Gutierrez-Lopez et al., 2016). These results demonstrated the power of cpDNA markers to distinguish cacao genotypes from one another.

In present study, chloroplast genomes sequencing of two cacao genotypes, DR-1 and Sca-12, were performed using an Illumina platform (Illumina Inc., USA). The study aimed to perform phylogenetic analysis of cacao genotypes and develop DNA barcode markers using cacao chloroplast sequences. The result presented here would be valuable to understand the structure of cacao's chloroplast genome as fundamental information useful for constructing the phylogenetic tree and for developing DNA barcoding as powerful tool in discriminating among cacao genotypes.

MATERIALS AND METHODS

Genetic materials and preparation of Chloroplast sequencing

Two cacao genotypes, DR-1 and Sca-12, analyzed in this study were derived from Research Institute for Industrial and Beverage Crops which is one of a scientific research center under the Indonesian Ministry of Agriculture located at Sukabumi, West Java Province, Indonesia. The DNA of cacao was isolated from healthy leaf tissue according to the DNA isolation procedure of CTAB (Allen et al., 2006). The extracted DNA were subjected to both quantity and quality assessment using Nanodrop ND-1000 (Thermo scientific). Afterwards, these two genotypes went through a sequencing process performed by Illumina sequencing platform (Illumina Inc., USA). The genomes of cacao chloroplast as well as 45S nuclear rDNA from two cacao genotypes were generated by *de novo* assembly method as previously mentioned by Kim et al. (2015a; 2015b). The quality of paired end reads was produced and subjected to assembly using CLC genomic workbench.

Annotation of cacao chloroplast genomes

Annotation of cacao chloroplast genomes were performed using the DOGMA program and corrected manually through BLAST search. The maps of cacao chloroplast genome were generated using OrganellarGenome DRAW tool (ORDRAW). Comparison of chloroplast sequences between two cacao genotypes sequenced in present study along with reference chloroplast genomes of cacao was accomplished using mVISTA program (Frazer et al., 2004). Furthermore, identification of repeat sequences was examined using REPuter program with the criteria as follows: sequence identity ≥90% and cutoff n ≥30 bp (Kurtz et al., 2001).

Phylogenetic analysis

A total of 12 cacao genotypes including two genotypes sequenced in this study (DR-1 and Sca-12) and ten genotypes previously sequenced (HQ336404, EET-64, ICS-01, ICS-06, ICS-39, Pentagonum, Sca-6, Stahel, Amelonado, and Criollo-22) were subjected for phylogenetic analysis. All the sequences of cacao chloroplast were then aligned with the MAFFT program. Dendrogram of 12 cacao genotypes was analyzed using Neighbor-joining method by a 1000 permutation in MEGA 6.0 (Tamura et al., 2013).

Exploitation of InDeL-based DNA barcode markers

The development of InDel barcode markers was carried out using the polymorphic regions found after alignment of the chloroplast sequences of DR-1 and Sca-12 genotypes. We used Primer 3 version 0.4.0 to design InDeL barcode primers. PCR analysis was conducted with a final volume of 25 µL which comprised 20 ng DNA, 2.5 mM dNTPs, 1X PCR buffer, 20 pmol each primer, and 2 units Taq DNA polymerase. The PCR condition was carried out according to the following criteria: denaturation for 5 minutes at 94 °C, 35 cycles of annealing process (30 s at 94 °C, 30 s at 54 °C, and 20 s at 72 °C) and final extension for 7 minutes at 72 °C. The amplification results were checked onto 1% agarose gel that had been stained using Gelred nucleic acid, then the gel was visualized using a gel documentation system.

RESULTS AND DISCUSSION

Characteristics of cacao chloroplast genome

We successfully sequenced chloroplast genome from two cacao genotypes, DR-1 and Sca-12 (Fig. 1) which resulted in total of 8,899,582 and 7,958,636 raw reads, respectively (Table 1). DR-1 genotype has similar with regards to chloroplast genome size to Sca-12 which ranged from 160,619 to160,649 bp in length. The typical of the chloroplast genomes of cacao genotypes were in quadripartite structure containing LSC, SSC and IR. Of which, LSC, IR, and SSC between respective genotypes have different length (Table 1).

Based on gene ontology (GO) annotation analysis, a total 112 genes were detected in cacao chloroplast genome which comprised 78 genes encoding protein, 30 genes of tRNA, and four genes of rRNA. Some detected genes (17 genes encoding protein, 11 tRNA, as well as four rRNAs) were present more than two copies in the genomes (Table 2). This annotation result revealed that gene content in the genome of cacao chloroplast is identical even in different species. Of which, gene content presented in cacao chloroplast genome is similar to that in *Artemisia* species (Liu et al., 2013). In addition, we also compared the IR/LSC and IR/SSC borders among two cacao genotypes (DR-1 and Sca-12) sequenced in present study with those among four cacao genotypes (Sca-6, Criollo-22, Stahel, HQ336404) which previously sequenced (Fig. 2). The results exhibited the gene structure among six cacao genotypes was relatively similar except for the gene presented between IR/LSC

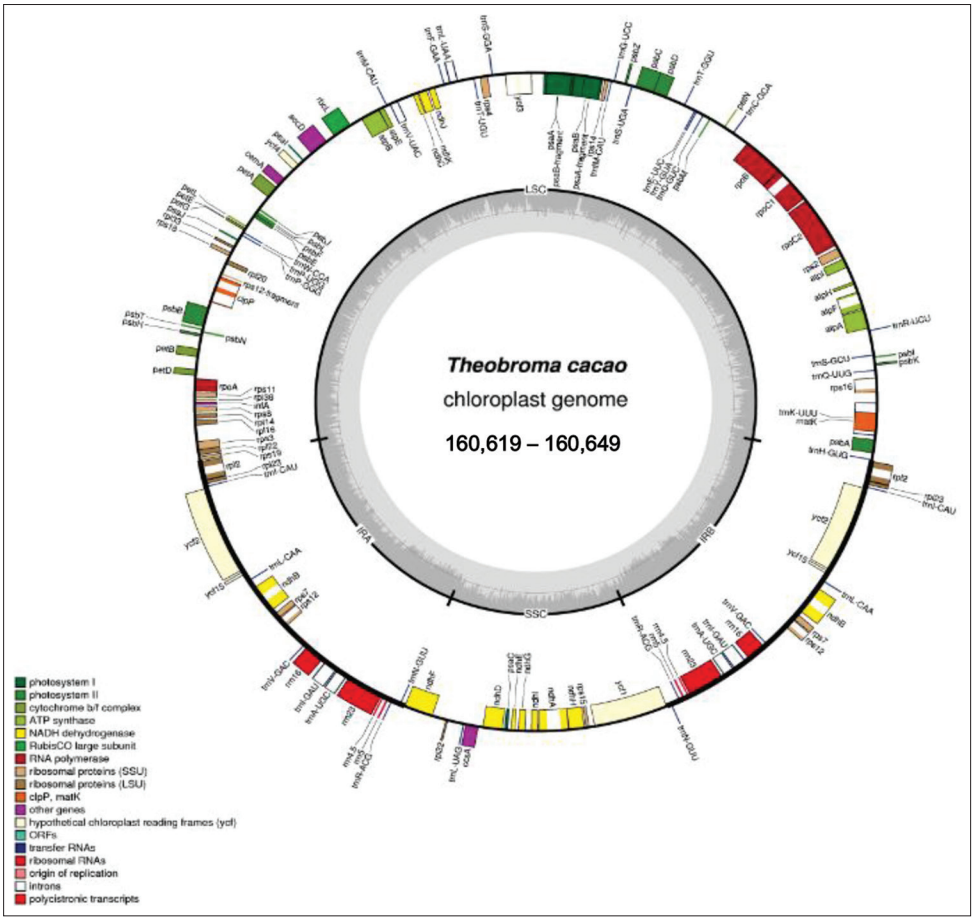

Fig 1. Map of *T. cacao* chloroplast genome. Transcription of genes is indicated in a clockwise (inside) and anticlockwise (outside) direction.

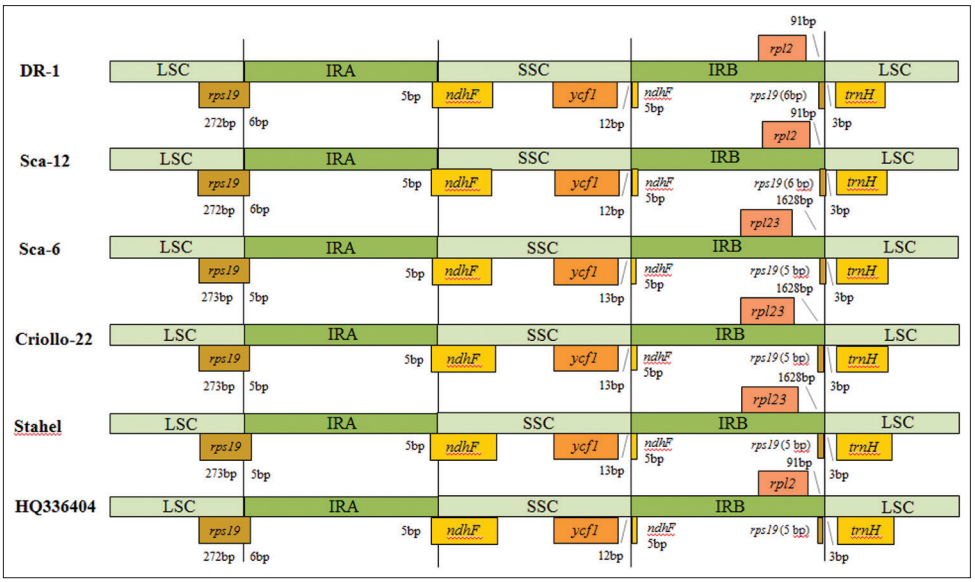

Fig 2. Characteristics of boundary regions found in SSC, LSC, and IR among 6 *T. cacao* chloroplast genomes.

Table 1: Features of two *Theobroma cacao* chloroplast genomes.

Parameters	DR-1	Sca-12
No. of raw read	8,899,582	7,958,636
Total read length (bp)	2,481,473,028	2,311,308,671
No. of mapped read	126,327	338,393
Chloroplast genome coverage (x)	183.56	508.62
Chloroplast genome size	160,649	160,619
Length of LSC (bp)	89,421	89,333
Length of IR (bp)	25,524	25,546
Length of SSC (bp)	20,180	20,194

LSC refers to large single copy region, SSC refers to small single copy region, IR refers to inverted repeat.

regions. Three genotypes (DR-1, Sca-12, and HQ336404) had *rpl2* gene between IR/LSC regions, whereas the other three genotypes (Sca-6, Criollo-22, and Stahel) possessed *rpl23* gene. Besides, there was also slightly differences in expansion of gene between IR/LSC and IR/SSC regions. The *rps19* gene overlapped the boundary region of LSC/IRA, resulting in *rps19* pseudogene fragments at about 5-6 bp in the IRB region. The other gene, called as *ndhF* spanned about 5 bp across IR/SSC border regions that resulted of ndhF pseudogene fragment in the IRB region. Expansion that occurred at four junctions of IR/SC regions can induce the length variation of cacao chloroplast genomes. This finding was similar to those found in *Epimedium* (Zhang et al., 2016) and *Veroniceae* (Choi et al., 2016).

Sequence variation among cacao genotypes

Sequence variation among two Cacao genotypes sequenced in current study and chloroplast genome of four cacao genotypes previously published were examined using mVISTA program. In this analysis, sequence annotation of *DR-1* genotypes was used as reference. The result showed that sequence variation was commonly found in non-coding area compared to coding area (Fig. 3), which proves that the coding region of chloroplast genome is more conserved. Nevertheless, we observed two variations occur in coding region of six cacao genomes, i.e. *rpoC2* and *ycf1*. The result showed that only low variation found among cacao genotypes, which indicated low mutation occured within cacao species. Interestingly, one coding sequences, *ycf1*, which showed variation in this study was also divergent in other crops (Liu et al., 2013; Kim et al., 2015a; Chen et al., 2015), indicating there have been high mutation rate occured in this region. The occurrence of sequence variation in coding region is largely caused by an indel mutation (Chen et al., 2015).

Variation among cacao genotypes was also observed from the presence of SNPs and indel in the chloroplast sequences. We compared the availability of SNP and InDel between 12 cacao genotypes. Of which, the highest number of SNP was found between Sca-12 vs DR-1 (67 SNPs). Meanwhile, the lowest number of SNP was observed between Stahel vs EET-64, Pentagonum vs ICS-39, and Criollo-22 vs Stahel (1 SNP). Interestingly, no SNP was found between Criollo-22 vs EET-64 genotypes (Table 3). In the case of indels, we observed that not all cacao genotypes contained indels. The different number of SNP and indel observed within the cacao genotypes could be due to the difference of nucleotide substitutions in the sequences of the same species. Another possible explanation is 12 cacao genotypes observed in this study undergo different mutation event that might be related to the different growing region. These sequence variations would be useful for studying genetic diversity among closely

Table 2: Genes variation presented in two cacao chloroplast genomes sequenced in present study

Gene category	Gene functions	Gene name
Self-replication	Ribosomal RNAs (rRNAs)	rrn16[a], 23[a], 4.5[a], 5[a]
	Transfer RNAs (tRNAs)	trnA-UGC[b], C-GCA, D-GUC, E-UUC, F-GAA, fM-CAU, G-UCC[a], H-GUG, I-CAU[a], I-GAU[b], K-UUU[a], L-CAA[a], L-UAA[a], L-UAG, M-CAU, N-GUU, P-UGG, P-GGG, Q-UUG, R-ACG[a], R-UCU, S-GCU, S-GGA, S-UGA, T-GGU, T-UGU, V-GAC[a], V-UAC[a], W-CCA, Y-GUA
	Ribosomal protein (small)	rps2, 3, 4, 7[a], 8, 11, 12[b], 14, 15, 16[a], 18, 19
	Ribosomal protein (large)	rpl2[b], 14, 16[a], 20, 23[a], 32, 33, 36
	DNA-dependent RNA polymerase	rpoA, B, C1[a], C2
Genes for photosynthesis	ATP synthase	atpA, B, E, F[a], H, I
	Cytochrome b6/f complex	petA, B[a], D[a], G, L, N
	NADH dehydrogenase	ndhA[a], B[b], C, D, E, F, G, H, I, J, K
	Photosystem I	psaA, B, C, I, J, ycf3[b], 4
	Photosystem II	psbA, B, C, D, E, F, H, I, J, K, L, M, N, T, Z
	Rubisco (Large subunit)	rbcL
Other genes	Maturase	matK
	Protease	clpP[b]
	Membrane protein envelope	cemA
	Subunit of Acetyl-CoA-carboxylase	accD
	c-type cytochrome synthesis gene	ccsA
Unknown function genes	ORF (ycf)	ycf1[a],2[a],5[a]

[a]presence in two copies [b]presence in more than three copies in cacao chloroplast genome

Table 3: The SNP and InDel identified among chloroplast genomes of 12 *T. cacao* genotypes

Species	DR1	Sca12	HQ336404	EET64	ICS01	ICS06	ICS39	Pentagonum	Sca6	Stahel	Amelonado	Criollo22
DR-1	/	41	29	41	41	41	41	36	41	42	41	43
Sca-12	67	/	27	6	6	6	6	6	6	6	6	6
HQ336404	60	46	/	24	24	24	24	24	24	24	24	24
EET-64	14	53	47	/	-	-	-	-	-	-	-	-
ICS-01	41	26	20	27	/	-	-	-	-	-	-	-
ICS-06	63	10	44	51	23	/	-	-	-	-	-	-
ICS-39	15	57	51	4	31	55	/	-	-	-	-	-
Pentagonum	16	58	51	5	32	56	1	/	-	-	-	-
Sca-6	63	2	44	47	24	8	55	55	/	-	-	-
Stahel	13	54	48	1	28	53	3	4	52	/	-	-
Amelonado	51	36	9	37	10	33	41	42	33	38	/	-
Criollo-22	14	53	47	-	27	51	4	5	52	1	37	/

related cacao species as well as for developing barcode marker to distinguish among cacao genotypes.

In present study, we also observed the number of repetitive sequences between 12 cacao genotypes which was analyzed using REPuter program (Fig. 4). Most of repeat sequences contained 30-39 bp in length. The result exhibited that each genotype possessed different number of repeats, i.e 18 and 19 repeats for DR-1 and Sca-12 genotypes, respectively (Table 4). Based on the repeat structure, these two cacao genotypes had three repeat sequences (i.e. forward, palindrome, and reverse). This finding is interesting, although the same species but showed different number and structure of repeat sequence, which might be due to an indel mutation event. Among four types of repeat sequence, forward and palindrome repeat are common in cacao species. These two repeats also found as common repeat in Epimedium (Zhang et al., 2016) and Veroniceae (Choi et al., 2016). According to previous studies, repeat sequences are helpful for analysis of genome rearrangement as well as for population genetic study and phylogenetic analysis because these repeat sequences could be a valuable source for marker development (Nie et al., 2012; Zhang et al., 2016).

Phylogenetic analysis using chloroplast sequences

The use of chloroplast sequences has brought a remarkable success in studying phylogenetic relatedness in many land plants (Liu et al., 2013; Kang et al., 2016; Lee et al., 2016a; Lee et al., 2016b). The utilization of chloroplast genome has become attractive for understanding phylogenetic relatedness among different plant species or within the same species because it has a simple genetic structure,

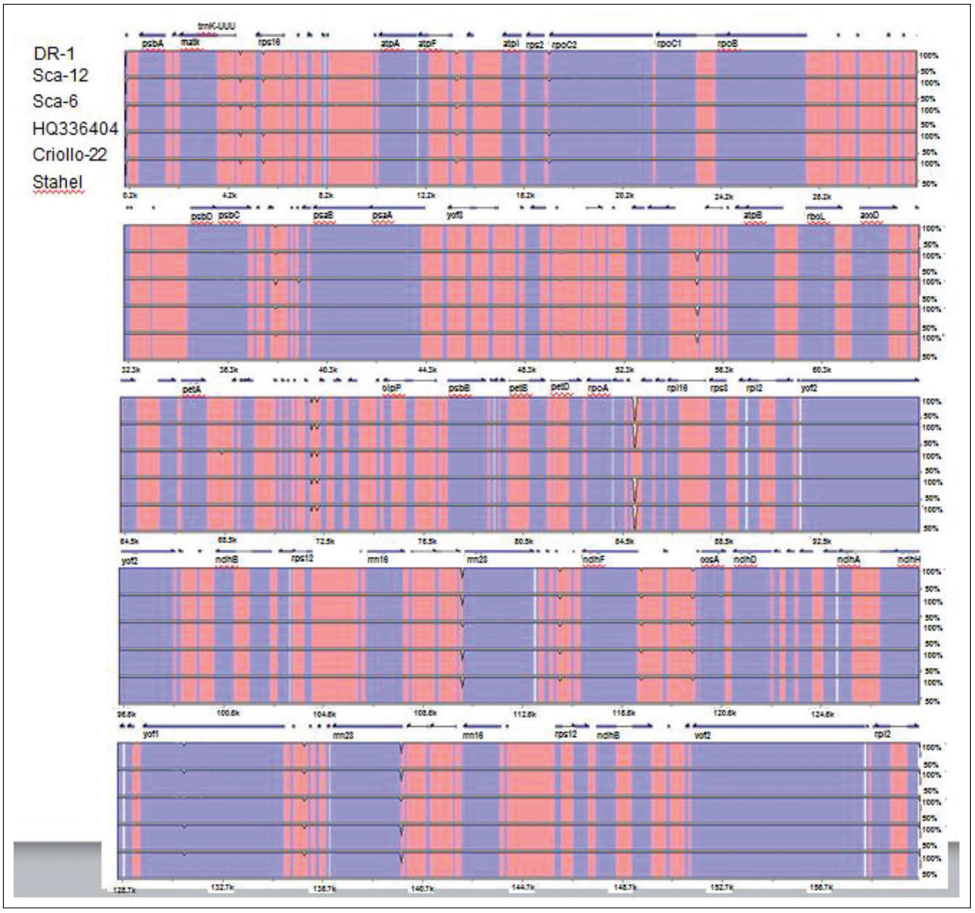

Fig 3. Chloroplast sequence variation in the six cacao genotypes analyzed through mVISTA.

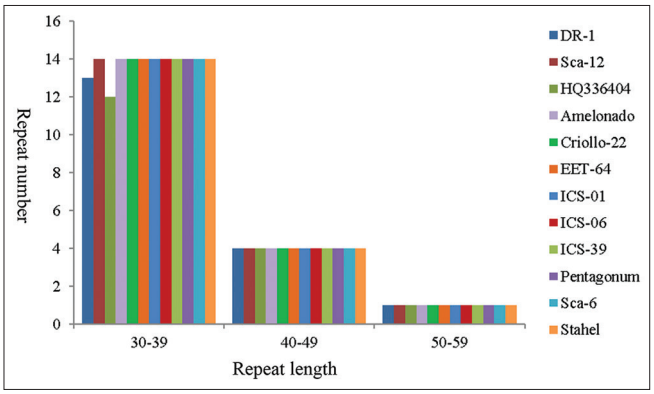

Fig 4. Repeat sequence frequency observed in chloroplast genomes of 12 *T. cacao* genotypes.

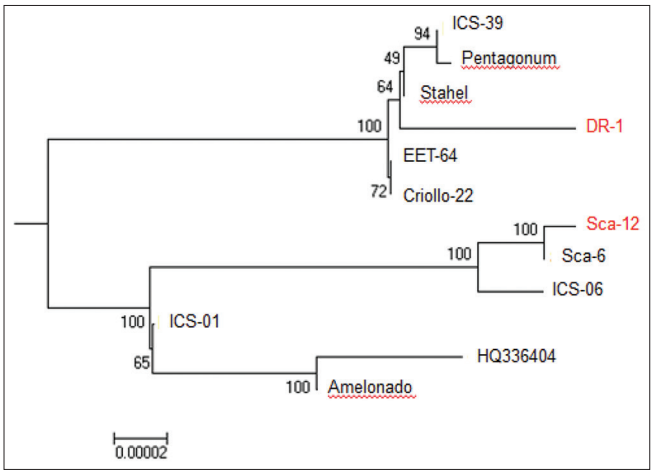

Fig 5. Phylogenetic tree of 12 *T. cacao* genotypes constructed using complete chloroplast genome sequence.

conserved, has high evolutionary rate, recombination rarely occurs, and commonly uniparental inheritance (Dong et al., 2012). Numerous studies have analyzed phylogenetic relationships based on multiple alignments of genes encoding proteins presented in the genomes of chloroplast in plant species (Chen et al., 2015; Zhang et al., 2016; Choi et al., 2016; Wang et al., 2016). Phylogenetic tree constructed from chloroplast sequences of 12 cacao genotypes used in present study was presented in Fig. 5. The resulting dendrogram demonstrated 12 genotypes of cacao clustered into 2 main groups. Members of the first cluster comprised six genotypes including ICS-39, Pentagonum, Stahel, DR-1, EET-64, and Criollo-22 and most of genotypes belonged to Criollo and Trinitario types. The second group also contained six genotypes including Sca-12, Sca-6, ICS-06, ICS-01, HQ336404, and Amelonado. All of genotypes member of group two belonged to the

Table 4: Characteristics of repetitive sequences and their locations in two *T. cacao* genotypes

No	Repeat length (bp)	Repeat type	First repeat			Second repeat		
			Position	Location	Region	Position	Location	Region
DR-1								
1	58	P	9914	*trnS-GCU - trnR-UCU*	LSC	9914	*trnS-GCU - trnR-UCU*	LSC
2	33	P	17006	*rps2 - rpoC2*	LSC	17006	*rps2 - rpoC2*	LSC
3	31	P	29191	*trnC-GCA – petN*	LSC	29191	*trnC-GCA – petN*	LSC
4	31	F	33323	*trnT-GGU – psbD*	LSC	33338	*trnT-GGU – psbD*	LSC
5	41	F	40983	*psaB*	LSC	43207	*psaA*	LSC
6	35	F	49698	*trnT-UGU - trnL-UAA*	LSC	49731	*trnT-UGU - trnL-UAA*	LSC
7	38	R	52253	*trnF-GAA – ndhJ*	LSC	52253	*trnF-GAA – ndhJ*	LSC
8	40	P	54361	*ndhC - trnV-UAC*	LSC	54361	*ndhC - trnV-UAC*	LSC
9	30	P	54658	*ndhC - trnV-UAC*	LSC	54658	*ndhC - trnV-UAC*	LSC
10	32	F	60793	*rbcL*	LSC	60831	*rbcL*	LSC
11	30	R	75421	*clpP*	LSC	75434	*clpP*	LSC
12	48	P	79277	*psbT – psbN*	LSC	79277	*psbT – psbN*	LSC
13	31	F	94257	*ycf2*	IRA	94278	*ycf2*	IRA
14	34	F	96715	*ycf2*	IRA	96751	*ycf2*	IRA
15	30	F	105040	*rps12 - trnV-GAC*	IRA	105067	*rps12 - trnV-GAC*	IRA
16	30	F	110179	*trnA-UGC - rrn23*	IRA	110184	*trnA-UGC - rrn23*	IRA
17	34	F	113283	*rrn4.5 - rrn5*	IRA	113315	*rrn4.5 - rrn5*	IRA
18	40	F	119362	*rpl32 - trnL-UAG*	SSC	119378	*rpl32 - trnL-UAG*	SSC
Sca-12								
1	58	P	9930	*trnS-GCU - trnR-UCU*	LSC	9930	*trnS-GCU - trnR-UCU*	LSC
2	33	P	17019	*rps2 - rpoC2*	LSC	17019	*rps2 - rpoC2*	LSC
3	31	P	29210	*trnC-GCA – petN*	LSC	29210	*trnC-GCA – petN*	LSC
4	31	F	33342	*trnT-GGU – psbD*	LSC	33357	*trnT-GGU – psbD*	LSC
5	41	F	41000	*psaB*	LSC	43224	*psaA*	LSC
6	35	F	49715	*trnT-UGU - trnL-UAA*	LSC	49748	*trnT-UGU - trnL-UAA*	LSC
7	38	R	52270	*trnF-GAA – ndhJ*	LSC	52270	*trnF-GAA – ndhJ*	LSC
8	40	P	54378	*ndhC - trnV-UAC*	LSC	54378	*ndhC - trnV-UAC*	LSC
9	30	P	54675	*ndhC - trnV-UAC*	LSC	54675	*ndhC - trnV-UAC*	IRA
10	32	F	60810	*rbcL*	LSC	60848	*rbcL – accD*	LSC
11	30	R	75415	*clpP*	LSC	75428	*clpP*	LSC
12	48	P	79270	*psbT – psbN*	LSC	79270	*psbT – psbN*	LSC
13	31	F	94169	*ycf2*	IRA	94190	*ycf2*	IRA
14	34	F	96627	*ycf2*	IRA	96663	*ycf2*	IRA
15	30	F	104952	*rps12 - trnV-GAC*	IRA	104979	*rps12 - trnV-GAC*	IRA
16	30	F	110090	*trnA-UGC - rrn23*	IRA	110125	*trnA-UGC - rrn23*	IRA
17	34	F	113224	*rrn4.5*	IRA	113256	*rrn4.5 - rrn5*	IRA
18	34	P	117299	*ndhF - rpl32*	SSC	117299	*ndhF - rpl32*	SSC
19	40	F	119299	*rpl32 - trnL-UAG*	SSC	119315	*rpl32 - trnL-UAG*	SSC

Forastero type. Of the six genotypes in the group one, ICS-39 exhibited closer relationship with Pentagonum, whereas EET-64 was placed as a sister to Criollo-22. In the group two, Sca-12 and Sca-6 had close relationship to each other, while HQ336404 showed closer relationship with Amelonado. These results suggested that chloroplast sequences are very useful for phylogenetic analysis because it can divide cacao genotypes based on their varietal group.

Development of indel-based barcode markers in cacao

We developed InDel-based barcode markers by comparative chloroplast genome analysis of two cacao genotypes, DR-1 and Sca-12. A total of three markers (Theca_indel01, Theca_indel02, and Theca_indel03) (Table 5) were designed using the polymorphic sites located at *trnA-UGC - rrn23, trnK-UUU-rps16*, and *rps16* intron regions (Fig. 6). All newly designed indel-based barcode markers are promisingly DNA markers potential to discriminate

Table 5: A set of indel-based barcode marker developed in present study

Marker	Forward and reverse sequences	Location	Product size (bp)
Theca_indel01	F: TGCACGATGCAATCAAAACA R: CGCCCATAAGCTTGTTGACTT	*trnK-UUU - rps16*	272/281
Theca_indel02	F: TTTTCTCCTCGTACGGCTCG R: AGGGGTTAGAGACCACTCAA	*rps16* intron	235/243
Theca_indel03	F: TTTACCCTGTGGCGGATGTC R: CACCGTAAGCCTTTCCTCGT	*trnA-UGC - rrn23*	252/282

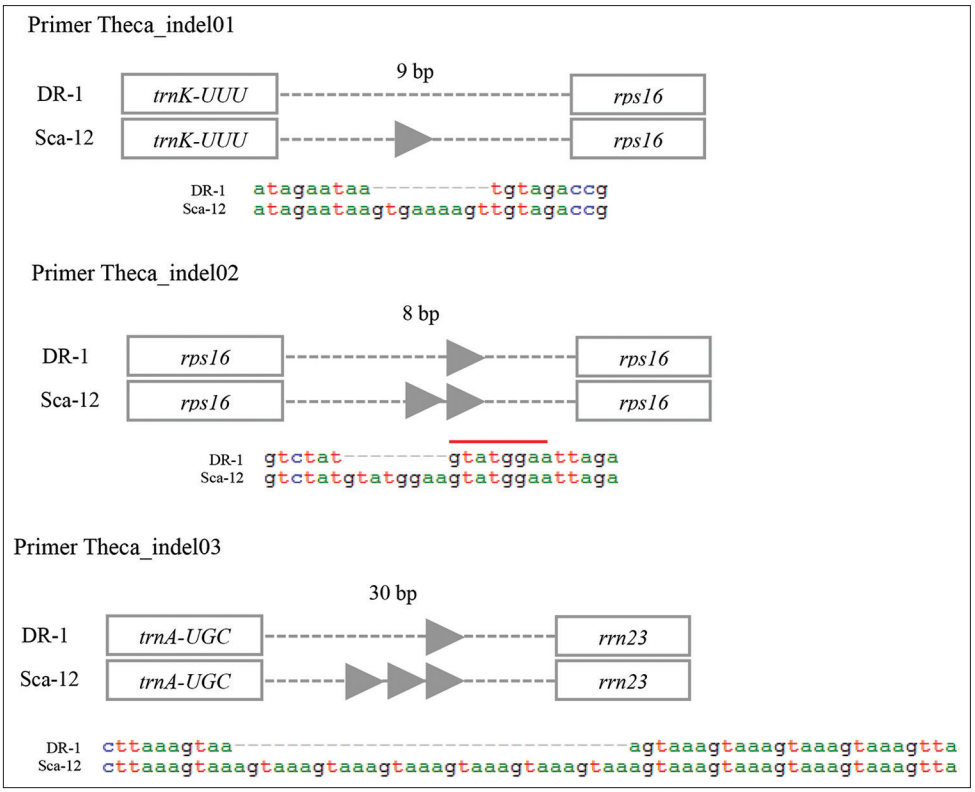

Fig 6. Polymorphic region used to design InDel-based barcode markers.

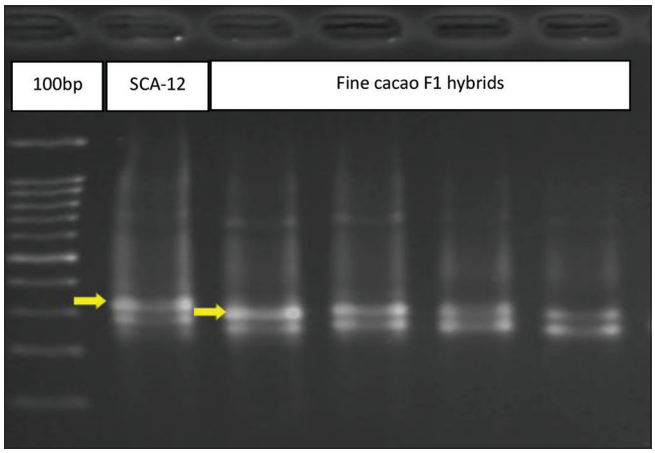

Fig 7. Primer Theca_indel03 successfully discriminate five cacao genotypes based on their varietal group.

cacao genotypes on the basis of their varietal group. By using these markers, we expect that Trinitario type can be discriminated from Forastero type.

We then used these three InDel-based barcode markers to differentiate five cacao genotypes, including one genotype of Forastero type (SCA-12) and four genotypes of Trinitario type (fine cacao F1 hybrids). The result showed that primer Theca_indel03 successfully distinguished Forastero and Trinitario types (Fig. 7). This result proved that the DNA barcode markers designed in this study would be beneficial for cultivar differentiation among cacao genotypes, which help breeders to select between Trinitario and Forastero type.

CONCLUSION

We have successfully developed chloroplast map of cacao with the size of about 160,619 bp to 160,649 bp. Gene annotation revealed that chloroplast genome of cacao plant contained 112 genes, comprised 78 genes encoding protein, 30 genes of tRNA, and four genes of rRNAs. Phylogenetic analysis constructed from the sequences of cacao chloroplast genome was successfully divided 12 cacao genotypes based on their varietal group (bulk and fine types). This result demonstrated the usefulness of cacao chloroplast sequences for constructing phylogenetic trees. Furthermore, using the REPuter program, we have identified 18-19 repetitive sequences as well as three repeat structures in cacao. Three indel-based DNA barcode markers have been developed using polymorphic sites located at *trnA-UGC - rrn23, trnK-UUU - rps16*, and *rps16* intron regions. One primer, Theca_indel03, successfully discriminated five cacao genotypes based on their varietal group. Overall, the result presented in this study revealed the diversity of chloroplast sequences in cacao that can be applied to create phylogenetic trees as well as design DNA barcode markers.

ACKNOWLEGEMENT

This study was supported by SMARTD program year 2016 under Indonesian Ministry of Agriculture. Authors would like to thank the head of Research Institute for Industrial and Beverage Crops, Indonesian Ministry of Agriculture and the Lab members of Functional Plants Laboratory, Department of Agriculture, Forestry, and Bioresources, College of Agriculture and Life Sciences, Seoul National University for providing laboratory facilities for molecular work activities.

Conflicts of interest

No conflict of interest is declared.

Author's contributions

NKI and R accomplished the research project in laboratory, collected the data, conducted statistical analyses, wrote and revised the manuscript. HSP supported laboratory work and data analyses. JYP supported laboratory work activity. TJY designed the project and edited the manuscript.

REFERENCES

Allen, G. C., M. A. Flores-Vergara, S. Krasynanski, S. Kumar and W. F. Thompson. 2006. A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. Nat. Prot. 1: 2320-2325.

Chen, J., Z. Hao, H. Xu, L. Yang, G. Liu, Y. Sheng, C. Zheng, W. Zheng, T. Cheng and J. Shi. 2015. The complete chloroplast genome sequence of the relict woody plant *Metasequoia glyptostroboides* Hu et Cheng. Front. Plant Sci. 6: 447.

Cheesman, E. E. 1944. Notes on the nomenclature, classification and possible relationships of cocoa populations. Trop. Agricu. 21: 144-159.

Choi, K. S., M. G. Chung and S. Park. 2016. The complete chloroplast genome sequences of three veroniceae species (*Plantaginaceae*): Comparative analysis and highly divergent regions. Front. Plant Sci. 7: 355.

Chumley, T. W., J. D. Palmer, J. P. Mower, H. M. Fourcade, P. J. Calie, J. L. Boore and R. K. Jansen. 2006. The complete chloroplast genome sequence of *Pelargonium* x *hortorum*: Organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Mol. Biol. Evol. 23: 2175-2190.

Daniell, H., C. S. Lin, M. Yu and W. J. Chang. 2016. Chloroplast genomes: Diversity, evolution, and applications in genetic engineering. Genome Biol. 17: 134.

Dong, W., J. Liu, J. Yu, L. Wang and S. Zhou. 2012. Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for dna barcoding. PLoS One. 7: e35071.

Frazer, K. A., L. Pachter, A. Poliakov, E. M. Rubin and I. Dubchak. 2004. VISTA: Computational tools for comparative genomics. Nucl. Acids Res. 32: W273-W279.

Gonzalez, M. A., C. Baraloto, J. Engel, S. A. Mori, P. Petronelli, B. Riéra, A. Roger, C. Thébaud and J. Chave. 2009. Identification of amazonian trees with DNA barcodes. PLoS One. 4: e7483.

Kane, N., S. Sveinsson, H. Dempewolf, J. Y. Yang, D. Zhang, J. M. M. Engels and Q. Cronk. 2012. Ultra-barcoding in cacao (*Theobroma* Spp.; *Malvaceae*) using whole chloroplast genomes and nuclear ribosomal DNA. Am. J. Bot. 99: 320-329.

Kang, S. H., K. Kim, J. H. Lee, B. O. Ahn, S. Y. Won, S. H. Sohn and J. S. Kim. 2016. The complete chloroplast genome sequence of medicinal plant, *Artemisia argyi*. Mitochondrial DNA Part B Resour. 1: 257-258.

Kim, K., S. C. Lee, J. Lee, H. O. Lee, H. J. Joh, N. H. Kim, H. S. Park and T. J. Yang. 2015a. Comprehensive survey of genetic diversity in chloroplast genomes and 45S nrDNAs within *Panax ginseng* species. PLoS One 10: e0117159.

Kim, K., S. C. Lee, J. Lee, Y. Yu, K. Yang, B. S. Choi, H. J. Koh, N. E. Waminal, H. I. Choi, N. H. Kim, W. Jang, H. S. Park, J. Lee, H. O. Lee, H. J. Joh, H. J. Lee, J. Y. Park, S. Perumal, M. Jayakodi, Y. S. Lee, B. Kim, D. Copetti, S. Kim, S. Kim, K. B. Lim, Y. D. Kim, J. Lee, K. S. Cho, B. S. Park, R. A. Wing and T. J. Yang. 2015b. Complete chloroplast and ribosomal sequences for 30 accessions elucidate evolution of *Oryza* AA genome species. Sci. Rep. 5: 15655.

Kurtz, S., J. V Choudhuri, E. Ohlebusch, C. Schleiermacher, J. Stoye and R. Giegerich. 2001. REPuter: The manifold applications of repeat analysis on a genomic scale. Nucl. Acids Res. 29: 4633-4642.

Lee, Y.S., S. Woo, J. K. Kim, J. Y. Park, N. K. Izzah, H. S. Park, J. H. Kang, T. J. Lee, S. H. Sung, K. B. Kang and T. J. Yang. 2022. Genetic and chemical markers for authentication of three *Artemisia* species: *A. capillaris*, *A. gmelinii*, and *A. fukudo*. PLoS One 17: e0264576.

Lee, Y. S., J. Y. Park, J. K. Kim, H. O. Lee, H. S. Park, S. C. Lee, J. H. Kang, T. J. Lee, S. H. Sung and T. J. Yang. 2016a. Complete chloroplast genome sequence of *Artemisia fukudo* Makino (*Asteraceae*). Mitochondrial DNA Part B Resour. 1: 376-377.

Lee, Y. S., J. Y. Park, J. K. Kim, H. O. Lee, H. S. Park, S. C. Lee, J. H. Kang, T. J. Lee, S. H. Sung and T. J. Yang. 2016b. The complete chloroplast genome sequences of *Artemisia gmelinii* and *Artemisia capillaris* (*Asteraceae*). Mitochondrial DNA Part B Resour. 1: 410-411.

Liu, Y., N. Huo, L. Dong, Y. Wang, S. Zhang, H. A. Young, X. Feng and Y. Q. Gu. 2013. Complete chloroplast genome sequences of Mongolia medicine *Artemisia frigida* and phylogenetic relationships with other plants. PLoS One 8: e57533.

Nock, C. J., D. L. E.Waters, M. A.Edwards, S. G. Bowen, N. Rice, G. M. Cordeiro and R. J. Henry. 2011. Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnol. J. 9: 328-333.

Ohyama, K., H. Fukuzawa, T. Kohchi, H. Shirai, T. Sano, S. Sano, K. Umesono, Y. Shiki, M. Takeuchi, Z. Chang, S. I. Aota, H. Inokuchi and H. Ozeki 1986. Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. Nature. 322: 572-574.

Seberg, O and G. Petersen. 2009. How many loci does it take to DNA barcode a crocus? PLoS One 4: e4598.

Shinozaki, K., M. Ohme, M. Tanaka, T. Wakasugi, N. Hayashida, T. Matsubayashi. 1986. The complete nucleotide sequence of the tobacco chloroplast genome: Its gene organization and expression. EMBO J. 5: 2043-2049.

Tamura, K., G. Stecher, D. Peterson, A. Filipski and S. Kumar. 2013. MEGA6: Molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30: 2725-2729.

Wang, Y., D. F. Zhan, X. Jia, W. L. Mei, H. F. Dai, X. T. Chen and S. Q. Peng. 2016. Complete chloroplast genome sequence of *Aquilaria sinensis* (Lour.) Gilg and evolution analysis within the Malvales order. Front. Plant Sci. 7:280.

Zhang, Y., L. Du, A. Liu, J. Chen, L. Wu, W. Hu, W. Hu, W. Zhang, K. Kim, S. C. Lee, T. J. Yang and Y. Wang. 2016. The complete chloroplast genome sequences of five *Epimedium* species: Lights into phylogenetic and taxonomic analyses. Front. Plant Sci. 7:306.