Journal of Biotechnology Letters
ISSN: 0976-7045 & E-ISSN: 0976-7053, Volume 3, Issue 1, 2012, pp.-28-31.
Available online at http://www.bioinfo.in/contents.php?id=62

ANALYSIS OF DIVERSITY OF NUCLEAR MICROSATELLITE OF *Coffea arabica* AS REVEALED BY DATA MINING

SENE K.H. AND ADMASSU B.

Ethiopian Institute of Agricultural Research (EIAR), Agricultural Biotechnology Laboratory, Holetta Agricultural Research Center, P.O. Box 2003, Addis Ababa, Ethiopia.
*Corresponding Author: Email- kagnewh2004@yahoo.com

Received: August 17, 2012; Accepted: September 20, 2012

Abstract- Novel molecular tools to develop microsatellite markers were mentioned elsewhere. Database screening before developing microsatellite markers was suggested. The prior knowledge of the microsatellite sequence has immense applications to redesign primers and isolate microsatellites by approaches of experimental biology. Due to a growing concern of molecular characterization of *C. arabica*, the number of microsatellite sequence deposited in genbank is increasing from time to time. Advancement in bioinformatics and computational genomics simplified retrieval of the nucleotide sequence of the microsatellite and the repeated motifs. In the present study, the microsatellite sequence was retrieved from genbank using the microsatellite locus/accession number. Microsatellite accessions collected from different published articles were used for the present attempt of SSR mining. The SSR sequence was examined. The efficiency of SSR mining is governed by several factors. All microsatellites of coffee could not be exhaustively explored. The compilation of the entire simple sequence repeat of *C. arabica* and sequence retrieval of the SSR from genbank will be meritorious for experimentalist to undertake molecular experiments using the sequence information of the existing SSR. The present study indicates the gap in SSR mining. And advancement of Coffee genomics will contribute to the compilation of all existing coffee SSR.
Keywords- Accession, *Coffea arabica*, Database, Genbank, Microsatellite, SSR Mining

Citation: Sene K.H. and Admassu B. (2012) Analysis of Diversity of Nuclear Microsatellite of *Coffea arabica* as Revealed by Data Mining. Journal of Biotechnology Letters, ISSN: 0976-7045 & E-ISSN: 0976-7053, Volume 3, Issue 1, pp.-28-31.

Introduction

The subgenus *Coffea,* which belongs to the family *Rubiaceae* consists of approximately 100 taxa [1]. Globally, the diversity within Coffea appears to be high [2]. The wild coffee plant (*C. arabica* L.) is indigenous to south western Ethiopia [3]. Central and west Africa, East Africa, Central Africa and Madagascar are the four groups of geographical origin in which coffee is organized [1]. Ethiopia is the origin and centre of diversity for tetraploid *C. arabica*, which is the only tetraploid coffee while the rest being diploid. Genetic diversity of coffee reduced when it get introduced from the place of its primary centre of diversity [4]. Hence, identification of Arabica coffee cultivar is crucial [3]. The genetic variation within and among Arabica coffee cultivars is limited [5]. Coffee is thought to be originated from a species, which looks like *Coffea eugenoids*. There exists gene flow from diploid coffee species to the tetraploid *C. arabica* [6]. SSR markers, biochemical markers like isozymes and morphological markers are some the tools applied in identification of coffee cultivars. The DNA based characterization, that is, application of SSR marker, remains to be the most promising methods to characterize coffee. It is crucial to study genetic diversity of coffee for further study of genetic improvement and conservation of germplasm [7]. The study of characterization research in the genus Coffea have ample of merits. Molecular characterization of *C. arabica* is of paramount importance for it has great contribution for resolving confusions in basic coffee taxonomy, monitoring diversity of coffee in time and space, conserving coffee germplasm, revisiting coffee accessions in genbank, filling former gaps in coffee sampling, monitoring molecular evolution and marker assisted selections. There are various repeat types in *C. arabica*. Some repeated nucleotides are CA, TG, GT, TC, AC, CT, GA, A in microsatellite of *C. arabica* [2]. Repeats could be mononucleotide dinucleotides, trinucleotides and others. SSR markers are commonly used in diversity study of coffee since they are known to be transferable within the coffee genus for six species, namely, *C. canephora*, *C. eugenioides*, *S. moore*, *C. heterocalyx*, *C. liberica*, *C. anthony* and *C. pseudozanguebariae* Bridson. Generally, SSR are transferable, co-dominant, polymorphic, multiallelic, informative and simpler. There existed several attempts in coffee SSR marker development and characterization over the last years. The diversity of 15 coffea species was studied via applying tools of 60 microsatellite markers. Two cultivated and two related wild coffees were compared [2]. Genetic diversity of *C. arabica* collected from Wollega, Illubabor, Keffa, Jimma and Sidamo were studied using 32 SSR markers [8]. SSR studies were conducted on *Coffea* species collected from north of Jimma, Kaffa, Sidamo, Dilla, Teppi-Mizan Teferi and Illubabor [9]. The origin of the cultivated *C. arabica L*. was studied using AFLP and SSR markers [4]. Eleven microsatellite loci of *C. arabica* were studied [10].

Although there existed several databases following the rapidly growing structural, functional and comparative genomics, the number of microsatellite sequence deposited in genbank are too limited. Also the number of databases related to coffee are quite few. Obviously, there are criteria to deposit a microsatellite sequence in genbank. The microsatellite must be sequenced and the clone sequence from EST must be available. In fact, sequenced microsatellites that are not submitted to genbank can't be accessed from NCBI server. Thus, there are a number of microsatellites, which are not sequenced and hence they can't be accessed from nucleotide databases. There may be chance that microsatellites sequence may be deposited in genbank, but there could be other peculiar key words or texts instead of accession number to retrieve the sequence. Probably, some authors may deposit microsatellite sequence in databases other than NCBI. Most importantly, several authors might have developed microsatellite. But, they may be on the virtue of submission of the sequence. Some authors might unable to send the microsatellite sequence for a couple of reasons. In the present paper, it is possible to address that compilation of the entire microsatellite is fundamental for conducting molecular experiments using tools of molecular markers. Availing list of microsatellites will aid to screen preferable satellites for experimental purpose. The current accessions used for mining the microsatellite sequence are obtained from variant published articles. In fact, articles couldn't be the only source of microsatellite accessions. Submission of microsatellite sequence without paper publication could be possible. The objective of this work is to address means of standardizing ways of retrieval microsatellite sequence, selecting proper key words/accession for sequence retrieval and validating merits of sequence retrieval. Obviously, organizing the microsatellite sequence in databases will aid in tracking the SSR of interest for a specific experiment.

Materials and Methods

At NCBI, the option "Nucleotide" was chosen from the lists and in the search space, the locus identifier /accession number was used to retrieve the microsatellite sequence. The option EST was also used. Accessions of microsatellite were collected from different published articles. A total of 154 microsatellite accessions were used in the present study to retrieve microsatellite of *C. arabica*. The published articles [2,9-11] are sources for the microsatellites selected in the present study. These microsatellites under study are AJ308753, AJ308755, AJ308774, AJ308779,AJ308782, AJ308790,AJ308809, AJ308837, AJ308838, CFGA2, CFGA35, CFGA38, CFGA54, CFGA55, CFGA69, CFGA74, CFGA75, CFGA91,CFGA92, CFGA99, CFGA100, CFGA189, CFGA202, CFGA207, CFGA227, CFGA236, CFGA249, CFGA276, CFGA280, CFGA281, CFGA285, CFGA311, CFGA465, CFGA485, CFGA491, CFGA494, CFGA499,CFGA502, CFGA529, CFGA547a, AJ250250, AJ25025l, AJ250252, AJ250253, AJ250254, AJ250255, AJ250256, AJ250257, A250258, AJ250259, AJ250260, CFGA574, CFGA627, CFGA792b, CFGA1122, CFGA1255, CFGA1258, CFCA14A, CFCA281, CFCA331, CFCA334, CFCA360, CFCA530, M20, M24, M25, M29, M32, M47, CarM028, CarM029, CarM030, CarM031, CarM032, CarM033, CarM034, CarM035, CarM036, CarM037, CarM038, CarM039, CarM040, CarM041, CarM042, CarM043, CarM044, CarM045, CarM046, CarM047, CarM048, CarM049, CarM050, CarM051, CarM052, CarM053, CarM054, CarM055, CarM056, CarM057, CarM058, CarM059, CarM060, CarM061, CarM062, CarM063, CarM064, CarM065, CarM066, CarM067, CarM068, CarM069, CarM070, CarM071, CarM072, CarM073, CarM074, CarM075, CarM076, CarM077, CarM078, CarM079, CarM080, CarM081, CarM082, CarM083, CarM084, CarM085, CarM086, CarM087, CarM088, CarM089, CarM090, CarM091, CarM092, CarM093, CarM094, CarM095, CarM096, CarM097, CarM098, CarM099, CarM100, CarM101, CarM102, CarM103, CarM104, CarM105, CarM106, CarM107, CarM108, Ccmp3, Ccmp6, Ccmp10 and NTCP8. Alternatively, the nucleotide sequences of the primers, which flank the microsatellite sequence, were used to search for the microsatellite at the database "Coffee DNA". Programs, softwares and editors like Vector NTI, AlignX, Clustal W/Clustal X, Bioedit, Genedoc, NJplot and Tree view were used for multiple sequence alignment, edition of microsatellite sequences and phylogeny analysis. Sequence analysis of simple sequence repeat to detect homologs was performed using BLASTN analysis. Number of repeats across accessions were plotted using excel spread sheet.

Results and Discussion

Most of the microsatellites of *C. arabica* are deposited in genbank. All of the profiles for genbank fulfilled. Both the repeated motif and the original sequence, which encompassed the motif, the original article and other details are displayed. Although the database "Coffee DNA" retrieved the microsatellite sequence, the existence of more than one microsatellite sequences complicated the search of the target microsatellite. In "Coffee DNA", the SeqMat option took the primer as an input sequence and one microsatellite may be retrieved based on this Primerblast tool. However, most of the primers pasted at the SeqMat resulted in more than one microsatellite. In this case, it is must to identify the target microsatellite. An attempt to align the forward and reverse primer back to the microsatellite screened the right microsatellite out of all the candidates. The use of primers as input sequence seems quite substantial. The satellites detected in *C. arabica* are repeats of 1, 2, 3 and 6 nucleotides. Some are perfect while others are imperfect. There existed both simple and compound repeat. Some representative examples of nucleotide repeats in *C. arabica* are CA, TG, GT, AG, TC, CT, TT, CG, CT/CA, TGA, AG/N/AG and CTCACA/CA. There are trends in variations in number of simple and perfect repeats microsatellites of *C. arabica* microsatellite. Some dinucleotide repeats like "AG" are common repeated motif. The number of repeats across accession number is shown in [Fig-1].

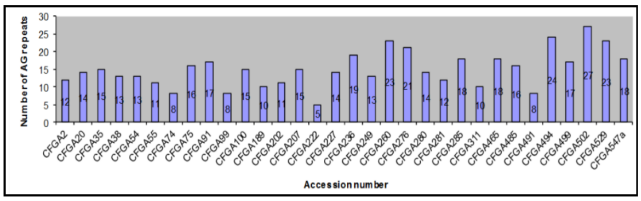

Fig. 1- Number of AG repeats across different accessions

The minimum number of repeats is 5, which is detected in microsatellite accession number "CFGA222" while the maximum number of repeats is 27, which is detected in CFGA502. The microsatellites of Coffee might have arisen through a number of factors. The coffee SSR may evolve through a number of factors. The

variation in "AG" repeat detected in the coffee genome is an exaggerated numeric, which reveals that "AG" repeats are frequent tandem. Such mechanisms of recombination as unequal crossing over or gene conversion and slippage during DNA replication may attribute to copy number fluctuation or expansion/contraction of microsatellite [12]. Some of the mined microsatellites are expressed sequenced tag-SSR. This is in line with previous work which stated presence of SSR in EST, that is, genic microsatellite, could be for lack of frameshift mutations in exons [13]. Accession numbers are not the only means of depositing microsatellite sequence. For example, we got hits with CFGA2. CFGA2 is not accession number. It is rather clone name. Rather AY102428 is the accession number for clone CFGA2. [Table-1] shows list of accession number and respective clone name.

Table 1- Clone name and corresponding accession number

S.No	Clone name	Accession number
1.	CFGA2	AY102428
2.	CFGA20	AY102429
3.	CFGA35	AY102430
4.	CFGA38	AY102431
5.	CFGA54	AY102432
6.	CFGA55	AY102433
7.	CFGA74	AY102435
8.	CFGA75	AY102436
9.	CFGA91	AY102437
10.	CFGA99	AY102439
11.	CFGA100	AY102440
12.	CFGA189	AY102441
13.	CFGA202	AY102442
14.	CFGA207	AY102443
15.	CFGA222	AY102444
16.	CFGA227	AY102445
17.	CFGA236	AY102446
18.	CFGA249	AY102447
19.	CFGA260	AY102448
20.	CFGA276	AY102449
21.	CFGA280	AY102450
22.	CFGA281	AY102451
23.	CFGA285	AY102452
24.	CFGA311	AY102453
25.	CFGA465	AY102454
26.	CFGA485	AY102455

This reveals that the SSR search efficiency will be affected for accession number is not the only code for depositing microsatellite sequence in databases. In case a given clone contained two accession numbers, it will be confusing. For example, we detected that accession numbers AJ308790 and AJ308796 belong to the same clone, namely, 12-4CTG. Clone name seems to lack specificity. We also detected that not all of the accessions of coffee have hits at NCBI. For instance, we can't get hits with CarM028, CarM029, CarM030, CarM031, CarM032, CarM033, CarM034, CarM035, CarM036, CarM037, CarM038, CarM039, CarM040, CarM041, CarM042, CarM043, CarM044, CarM045, CarM046, CarM047, CarM048, CarM049, CarM050, CarM051, CarM052, CarM053, CarM054, CarM055, CarM056, CarM057, CarM058, CarM059, CarM060, CarM061, CarM062, CarM063, CarM064, CarM065, CarM066, CarM067, CarM068, CarM069, CarM070, CarM071, CarM072, CarM073, CarM074, CarM075, CarM076, CarM077, CarM078, CarM079, CarM080, CarM081, CarM082, CarM083, CarM084, CarM085, CarM086, CarM087, CarM088, CarM089, CarM090, CarM091, CarM092, CarM093, CarM094, CarM095, CarM096, CarM097, CarM098, CarM099, CarM100, CarM101, CarM102, CarM103, CarM104, CarM105, CarM106, CarM107, CarM108 and NTCP8. As aforementioned, some accessions like Ccmp3, Ccmp6, Ccmp10, M20 and M24 were published to be microsatellite locus for coffee [14]. But, at NCBI, these locus tags/ accessions gave hit for other organisms, not for coffee. The BLAST analysis using a query sequence of AJ308755 is shown in [Table-2]. There existed four homologs. These are all microsatellites of coffee. Such effort of homolog scanning is also good to know more homologs for selecting candidate SSR for experiment. But, in these homologs no SSR is conserved revealing that their homology is not due to the SSR motif. They are homologues for some other sequences. The sequence alignment and phylogeny of these homologs is shown in [Fig-2] and [Fig-3]. There could be possibilities of maximizing SSR mining via BLAST analysis. However, not all of the SSR detect more homologs. For example, CFGA20 gave hit for itself, CFGA20 only. The same is true to CFGA207. CFGA465 gave hits for itself and other one homology, CFGA547a and CFGA465. CFGA547a also detected itself and CFGA465. In fact other SSR gave hits of more than two. AJ308755 gave four hits. Published microsatellites may not be deposited in genbank. CFGA465 was published to be "AG" repeats [9]. CFGA465 is a "CT" repeat in genbank. It means that CFGA465 deposited in genbank is different from the published CFGA465 regardless of code overlap.]

Table 2- BLASTN analysis

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
AJ308755.1	*Coffea arabica* microsatellite DNA, clone 34-2CTG	416	416	100%	3e-113	100%
AJ250258.1	*Coffea arabica* microsatellite DNA, clone ZapII.32	416	416	100%	3e-113	100%
AJ308796.1	*Coffea arabica* microsatellite DNA, clone 12-4CTG	268	268	67%	9e-69	99%
AJ308790.1	*Coffea arabica* microsatellite DNA, clone 12-4CTG	265	265	67%	1e-67	98%

Conclusion and Recommendations

It is recommendable that exploration of all the databases is the sole option to avail the entire microsatellite loci of *C. arabica*. At least comparison of two databases may signify existing gaps. Exploring other databases will enhance efficiency of SSR mining. The detection of other such databases as MoccaDB and TropGene indicates that more databases specific to coffee might exist. MoccaDB, a rubiaceae database searches SSR by species like *C.arabica*. It is also time to speculate that literature mining must be exhaustive enough. The knowledge of the overall SSR of *C. arabica* and their respective sequence will aid experimentalists to connect insilico biology to real life laboratory based experiments. Some databases may seem generic. For instance, using the primer sequence as a query sequence, the database "Coffee DNA" (a database for Coffee Genomics) retrieved more microsatellites that are not specific to that primer. Unlike coffee DNA, NCBI is very specific for it gave hits to that specific accession number only. However, the detection of more number of satellites as per single primer could be homolog detection. And this has a merit of mining more satellites. Homologs can be detected at NCBI if and only if

BLAST search is used. Indeed there is a need of standardization of the means of SSR mining and sequence retrieval to maximize mining of all the microsatellites developed for coffee. In the present study, no redundant deposition of sequence was detected. That also indicates that sequenced microsatellites were not deposited given that they were already deposited for the first time. Repeat number changes following microsatellite evolution will also affect the number of existing SSR for even a single extra SSR change can alter the microsatellite sequence. Future deposition of sequences will consider both newly developed coffee SSR and formerly developed SSR as long as there is change in number of repeats in the motif. A future attempt to expand the number of databases specific to coffee SSR or having a single mega database that considers all of the coffee SSR will be the way to go in order to access microsatellite of coffee with ease.

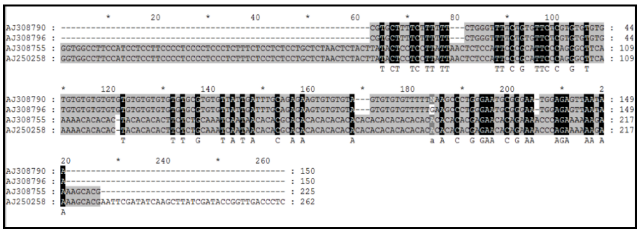

Fig. 2- Multiple sequence alignment of homologs of AJ308755

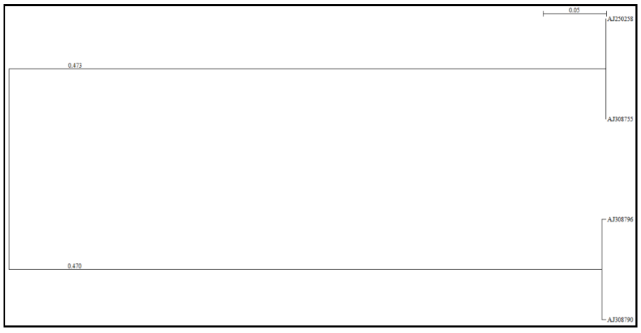

Fig. 3- Phylogeny of homologs of AJ308755

Acknowledgment

We thank you Agricultural Biotechnology laboratory of Ethiopia for availing computers with internet facility.

References

[1] Cros J., Lashermes P., Marmey P., Anthony F., Hamon S. and Charrier A. (1994) *Biotechnologie.*

[2] Cubry P., Musali P., Legnate H., Pot D., Bellis F.D., Poncet V., Anthony F., Dufour M. and Leroy T. (2008) *Genome*, 51.

[3] Sera T., Ruas P.M., Ruas C.D.F., Diniz L.E.C., Carvalho V.D.P., Rampim L., Ruas E.A. and Silveira S.R.D (2003) *Genetics and Molecular Biology,* 26.

[4] Anthony F., Combes M.C., Astorga C., Bertrand B., Graziosi G. and Lashermes P. (2002) *Theor. Appl. Genet.*, 104.

[5] Steiger D.L., H.N.C.M.P., W.M.C. and R.O.R.V.A.M. (2002) *Theor. Appl. Genet.*, 105.

[6] Herrera C.J., Marie Combes C., Cortina H. and Lashermes P. (2004)*Genome*, 47.

[7] Dessalegn Y., Herselman L. and Labuschagne2 M.T (2008) *African Journal of Biotechnology*, 7.

[8] Alemayehu N.T. (2007) *Ethiopian Journal of Applied Sciences and Technology,* 1(1).

[9] Moncada P. and McCouch S. (2004) *Genome*, 47.

[10] Combes M.C., Andrzejewski S., Anthony F., Bertrand B., Rovelli P., Graziosis G. and Lashermes P. (2000) *Molecular Ecology,* 9.

[11] Cristancho M. and Gaitán Á. (2008) *Crop Breeding and Applied Biotechnology,* 8.

[12] Hancock J.M. and Simon M. (2005) *Gene,* 345.

[13] Rajeev V.K., Andreas G. and Mark S.E. (2005) *Trends in Biotechnology,* 23.

[14] Vieira E.S.N., Pinho E.V.D.R.V., Carvalho M.G.G., Esselink D.G. and Vosman B. (2010) *Genetics and Molecular Biology*, 33(3).