International Journal of Computational Bioinformatics and In Silico Modeling
Vol. 5, No. 3 (2016): 815-818
Research Article
Open Access

ISSN: 2320-0634

In Silico Analysis of Genome wide microsatellite DNA marker in Coffee (*Coffea arabica* L.)

Virupaksh U. Patil[1]*, Vanishree, G[1], Hemant B. Kardile[1], Vikas Jindal[2], SK Dutta,[3] KK Chaturvedi[4] and Chakrabarti, S.K[1]

[1] *ICAR-Central Potato Research Institute, Shimla-171001 (Himachal Pradesh), India.*
[2] *Punjab Agriculture University, Ludhiana-141 004 (Punjab), India.*
[3] *ICAR-Research Complex for NEH Region, Mizoram Centre, Kolasib-796081, (Mizoram), India.*
[4] *ICAR-Indian Agriculture Statistical Research Institute, Pusa, New Delhi-110012 (Delhi), India.*

*Corresponding author: Virupaksh U. Patil; email: veerubt@gmail.com

Received: 31 March 2016 Accepted: 19 April 2016 Online: 02 May 2016

ABSTRACT

Microsatellites, also known as simple sequence repeats (SSRs) or short tandem repeats (STRs), are repeating sequences of 2-6 base pairs of DNA known for the ubiquitous existence in both prokaryotic and eukaryotic genomes. These markers are crucial for the marker assisted breeding of any crops. Difficulties in coffee breeding and wanting of new cultivars resistance to diseases and pest has urged the global coffee breeders to complement the traditional breeding with molecula markers assisted breeding. The present study reveals genome-wide presence and distribution of the microsatellite markers in the coffee. A total of 25,574 protein coding genes are predicted from coffee 3,34,404 SSRs were present in the genome of which 12.1% in the compound form. SSRs present in the gene sequence were 45,115, but only 2150 of these were actual genic SSRs. The information genome-wide microsatellites in coffee and specially the genic SSRs can be used in coffee breeding.

Keywords: Coffee, Genome-wide, Microsatellite, Marker, *in silico*.

1. INTRODUCTION

Coffee is the most important commercial commodity in the international agricultural trade representing a significant source of income to nearly 80 million people especially in the countries of Africa, Asia and South America. Its total global market values around 173.4 billion US$ and is increasing steadily with 1.9% annually. Coffee has become a fashionable drink over the last decade proven by the fact that coffee drink, so long most limited to cafes, is making a fresh charge into urban households in most of the developing countries including India. Over the last 50 years, there has been steady growth in the world coffee production. In the year 2012/13, world coffee production reached 145.1 million bags, the largest on record. South America continues to be the world's leading region with annual production 52.5 million bags and Brazil tops with 35.8 million bags [1]. The total coffee production of the world is 10.1 million tones with a productivity of 8792 kg/ha [2]. Of the 20 leading exporting countries, Vietnam recorded a highest growth rate in its exports (13.9%), followed by Peru (6.2%), Nicaragua (7.3%), Honduras (5.2%), and India (5.7%). Countries not only earn by coffee export but also by re-exporting its processed products. The re-export of coffee alone values US$ 14.7 billion globally [1]. The coffee belongs to the *rubiaceae* family which has 11,000 species distributed in 660 genus and is the fourth largest family in angeosperms (flowering plants). It is a brewed beverage prepared from the roasted or backed seeds of several species of an evergreen shrub of the genus *Coffea* originated from inter-tropical regions of Africa and Madagascar [3]. Although the *Coffea* genus includes more than 124 species, commercial coffee production relies exclusively on two related species, *Coffea arabica* (2n=4x=44) is a recent allotetraploid derived from

spontaneous hybridization between *Coffea canephora* and *Coffea eugenioides* and covers around 75 to 80% of the 11 million hectares and *Coffea canephora* (2n=2x-22) is an out cross diploid consisting of polymorphic populations of strongly heterozygous individuals and covers around 20 to 25% of the total cultivated area around the globe (http://www.ico.org) [4].

Coffee production in India and world is challenged by several diseases and pests, one of the most devastating disease is leaf rust which caused havoc in Central American countries in 2012/13. Total damage was estimated at around 2.7 million bags, costing a total of US$500 million, apart from its economic loss it also caused social losses i.e. it was estimated that 374,000 jobs were lost in the season [1]. So, the researchers in today's world are looking to produce crops that possess desirable characteristics, such as high yields, resistance to disease and many other characteristics that will benefit the crop in long term. The never ending advances in sequencing technologies provided opportunities to target not only the model plant species with small genome sizes, but many cultivated and economically important plant species like coffee for sequencing, identifying millions of novel markers, agronomically important genes, knowledge of which can directly translated into crop improvement. The knowledge of genetic structure based on molecular markers such as microsatellites [5] plays important role in population genetic studies, gene regulation and genome evolution [6]. These have proved useful in marker assisted selection of desirable traits to which they are linked, hence are the markers of choice for genome mapping studies. Advances in genome sequencing have contributed greatly to biological science. Till date, many crop genomes have been sequenced (www.ensembl.org/info/about/species.html), coffee being the very recently sequenced genome [7]. There is need to develop platform for mining genic microsatellites to ensure their better utilization as molecular markers, their abundance, distribution, evolution and putative function, if any. An SSR (Simple Sequence Repeat) consists of a variable number of tandem repeats of a 1 to 5 nucleotide motif. The most frequently observed repeat in plant species is $(AT)_n$ which has been observed to be randomly distributed across the genome. This length polymorphism is detected by PCR amplification of genomic DNA with specific primers flanking the SSR containing region. Among DNA markers, microsatellites have been chosen over other markers because of their simplicity, ubiquity, distribution across the genome, co-dominant behavior, multi-allelism, reproducibility, somatically stable, easy to assay using PCR and high level of polymorphism detected [8]. Moreover, although SSRs represent hypervariable areas of the genome, they are sufficiently conserved to be inherited for several generations in a mendelian fashion [9]. Till date there is no thorough *in silico* STR (Sequence Tagged Repeats) marker mining from coffee genome to represent more holistic and cumulative variability of genome to be used in gene pool or biodiversity analysis and gene/QTL mapping. Here, we present mining and analysis of STRs in coffee, genome wide as well as chromosome wise.

2. MATERIALS AND METHODS

The chromosome wise coffee genome, available in public domain (http://www.coffee-genome.org/download) [7, 10] was downloaded in FASTA format. All the 11 available chromosomes of the genome and the sequences not assigned to any chromosomes (chr0) were sliced into convenient sizes using PERL script for feeding into MIcroSAtellite identification (MISA) tool (http://pgrc.ipk-gaterleben.de/misa). The output data file was used for the analysis for getting an overview and to characterize the genome-wide distribution of microsatellite . It was observed that 87.9% of the total STRs were of simple/perfect in nature, whereas a 12.1% of the total 2,66,780 microsatellites were of compound in nature (includes both simple and interrupted) type, respectively. The 'single nucleotide' repeat type (mono-mer) was found to occur dominantly followed by 'di', tri, tetra, penta and hexa-mer was the least abundant type in the coffee genome (Figure 1). This phenomenon of STRs has been proved even in rice where the results from screening of genomic library suggest that there are about 5,700-10,000 microsatellites, with the relative frequency of different repeats decreasing with increasing size of the motif (McCouch *et al*. 1997).

3. RESULTS AND DISCUSSION

A total of 54% STRs had highest GC content (30 to 40%), whereas only 0.01% of the total STRs possessed <10% GC content. When the microsatellites were mapped on to individual chromosomes of the coffee genome, it was observed that chromosome 2 possessed the highest number of STRs followed by chromosome 6. In general, the number of STR markers present on the chromosome found to be directly correlated with the size of the chromosome. The biological and evolutionary significance of microsatellites is still not well understood. This may be probably because of the involvement of trinucleotide repeats in the recombination event, the larger chromosome size requires greater the number of repeats to have higher possibilities of cross over at each point on the chromosome [11]. Table 1 shows the distribution of simple and compound STRs along with repeat types (mono, di, tri, tetra, penta and hexa) on each chromosome.

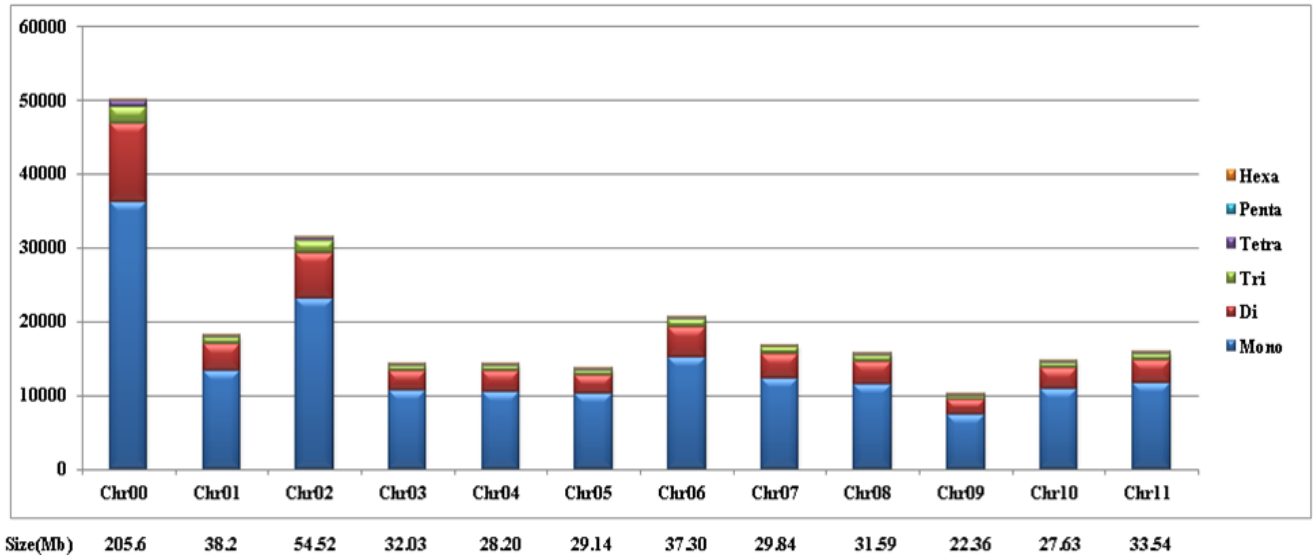

Figure 1: Graphical representation of the type and distribution of microsatellites on chromosomes of coffee.

Table 1: Chromosome-wise distribution of STRs based on the motif type in coffee.

Chr	Simple						Compound
	Mono	Di	Tri	Tetra	Penta	Hexa	
Chr00*	36145	10612	2281	819	125	81	5859
Chr01	13354	3578	944	170	54	32	2618
Chr02	23147	6190	1623	253	59	32	4585
Chr03	10712	2560	720	140	28	21	1944
Chr04	10497	2732	712	122	29	22	1985
Chr05	10161	2507	647	136	28	13	1932
Chr06	15153	4058	1098	189	46	24	3037
Chr07	12333	3284	855	136	30	21	2473
Chr08	11536	2969	847	153	34	18	2167
Chr09	7341	2012	514	96	31	8	1414
Chr10	10847	2812	690	136	32	21	2012
Chr11	11612	3154	803	166	42	17	2350
Total	172838	46468	11734	2516	538	310	32376

*Chr00 includes all the contigs which are not anchored on to any of the coffee chromosomes yet.

These STRs, when classified based on the size, 45.6% and 44.0% were found to be less than 10 bp and between 10-15 bp length respectively, whereas, only 0.8% were found to be having length more than 25bp. There are 25,574 predicted genes in coffee which contain 45,113 STRs in them (genic SSRs) of these 42,963 (95.2%) present in the UTR (Un-Translated Region) and only 2,150 (4.8%) are present on the coding region which are called genic SSRs. A total of 1,851 genes in the coffee genome contain the genic SSRs which may directly exploited for the associating with the biotic or abiotic stresses and can be used in breeding programs to develop high yielding and disease resistant coffee cultivars. Motif A/T was found to be more abundant than C/G in exons in all the taxa studied by [12], which is in agreement with our data. It remains unknown why certain repeat motifs are more common than others, or the reason they vary so much among or even within taxa. Furthermore, SSR motifs, abundance, and mutation rates are different among species, with a wide range of genetic properties [13].

4. CONCLUSION

On the basis of these findings and the previous data from other authors, we can conclude that there is a good potential for using the present approach for the targeted isolation of single or multiple, physically clustered SSRs linked to any *Coffee* gene that has been mapped using DNA-based markers. Further mining within the available databases will be needed if unique primer pairs for *Coffea* spp. are requested for genetic discrimination.

5. ACKNOWLEDGEMENTS

Authors thankfully acknowledge to Dr. Sanjeev Kumar, IASR, New Delhi for his valuable guidance and research input

6. REFERENCES

1. International Coffee Council (ICC, 2014) World coffee trade (1963-2013) A review of the markets, challenges, and opportunities facing the sector. 24, Feb 2014 (France).
2. FAOSTAT, 2013. http://faostat.fao.org/ as accessed on 29 December, 2014.

3. Davis AP, Govaerts R, Bridson DM. et al. (2006). An annotated taxonomic conspectus of the genus *Coffea* (*Rubiaceae*). *Bot. J. Linnean. Soc.*, 152: 465–512.
4. Maurin O, Davis AP, Chester M, et al. (2007). Towards a Phylogeny for *Coffea* (Rubiaceae): Identifying well-supported lineages based on nuclear and plastid DNA sequences. *Ann. Bot*., 100: 1565–1583.
5. Gonçalves-Vidigal MC and Rubiano LB. (2011) Development and application of microsatellites in plant breeding. *Crop Breed. Appl. Biotechs*., **S1**, 66-72.
6. Kashi Y and King DG. (2006). Simple sequence repeats as advantageous mutators in evolution. *Trends Genet*., 22: 253–259.
7. Denoeud F, Carretero-Paulet L, Dereeper et al. (2014). The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science*, 345: 1181-1184.
8. Al-Murish TM, Elshafei AA, et al. (2013). Genetic diversity of coffee (*Coffea arabica* L.) in Yemen via SRAP, TRAP and SSR markers. *J Agric Food Chem*., 11(2): 411-416.
9. Morgante M and Olivieri AM (1993). PCR-amplified microsatellite as markers in plant genetics. *Pl. J*., 3: 175-182.
10. Dereeper A, Bocs S, Rouard M. et al. (2014). The coffee genome hub: a resource for coffee genomes. *Nucl. Acid* Res. doi: 10.1093/nar/gku1108
11. Guo WJ, Ling J and Li P. (2009). Consensus features of microsatellite distribution: Microsatellite contents are universally correlated with recombination rates and are preferentially depressed by centromeres in multicellular eukaryotic genomes. *Genomics*, 93: 323-331.
12. Tóth G, Gáspari Z and Jurka J (2000). Microsatellites in different eukaryotic genomes: Survey and analysis. *Genome Res* 10: 967-981.
13. Cruz F, Pérez M and Presa P (2005). Distribution and abundance of microsatellites in the genome of bivalves. *Gene* 346: 241-247.

*****