IDENTIFICAÇÃO E CARACTERIZAÇÃO DE MICROSSATÉLITES DE *COFFEA ARABICA* A PARTIR DE DADOS DE SEQUENCIAMENTO DE RNA E DE BACS

Bruna Silvestre Rodrigues da Silva[1]; Sandra Bellodi Cação[2]; Suzana Tiemi Ivamoto[3]; Juliana Costa Silva[4]; Douglas Silva Domingues[5]; Luiz Filipe Protasio Pereira[6]

[1]Bolsista Consórcio Pesquisa Café, Londrina – PR, brunasilvestrerodrigues@hotmail.com
[2]Pesquisador, Dr, Agronomia – UEL, Londrina – PR, sandracacao@sercomtel.com.br
[3]Doutoranda de Genética e Biologia Molecular – UEL, Londrina – PR, suzanatiemi@yahoo.com.br
[4]Técnico Bolsista de Bioinformática, Iapar, Londrina – PR, jujucostasilvaanalise@gmail.com
[5]Pesquisador, Dr, Instituto Agronômico do Paraná , Londrina - PR, doug@iapar.br
[6]Pesquisador, Dr, Embrapa Café, Brasília – DF, filipe.pereira@embrapa.br

RESUMO: O café é uma das principais commodities agrícolas mundiais, sendo consumido regularmente por 40% de toda população. As espécies, *Coffea arabica L.* e *Coffea canephora P.* são as de maior importância respondendo por 65% e 35% da produção mundial, respectivamente. Em *C. arabica*, o desenvolvimento de novas cultivares é demorado, podendo levar mais de 20 anos para finalização dos trabalhos. Além disso, estreita base genética de *C. arabica* dificulta a obtenção de cultivares resistentes a várias pragas e doenças, assim como uma maior tolerância a estresses abióticos. A utilização de marcadores moleculares para análise da diversidade e seleção de genótipos representa uma importante ferramenta para auxiliar o melhoramento e tentar diminuir esses problemas. Neste trabalho foi realizada identificação e análise *in silico* de SSR's a partir de dados de RNA-seq de Iapar 59 e de sequências de BACs de Hibrido de Timor 832/2 (CaHT). Para Iapar 59 foram analisados 77 contigs, encontrados 39 SSR's e desenhados 21 pares de oligonucleotídeos. Para CaHT foram analisados também 77 contigs, encontrados 35 SSR's e desenhados 29 pares de oligonucleotídeos. Os motivos com maior frequência foram di, tri e tetranucleotídeos, sendo $(AT)_n$ o motivo mais frequente entre os dois genótipos. Esta busca de sequências repetitivas é de grande importância para futura validação e aumento desses marcadores para estudos de diversidade genética, mapeamento, e associação com características agronômicas de interesse.

PALAVRAS-CHAVE: marcadores SSR's, análise *in silico*, RNA-seq, cromossomo artificial de bactéria.

IDENTIFICATION AND CHARACTERIZATION OF COFFEA ARABICA MICROSATELLITE FROM RNA SEQUENCING DATE AND BACS

ABSTRACT: Coffee is a main agricutural commodity worldwide and is regularly consumed by 40% of the entire population. The species *Coffea arabica L.* and *Coffea canephora P.* are the most important, responsible for 65% and 35% of the world production, respectively. The development of new cultivars of *C. Arabica* is prolonged, can take up to 20 years to finally complete the work. Furthermore, the narrow genetic base of C. Arabica is a bottle neck to produce cultivars with resistance to various pests and diseases, thus as greater tolerance to abiotic stresses. The use of molecular markers for analysis of diversity and selection of genotypes can be an important tool for the breeding programs in order to try to reduce these problems. In this work we started the *in silico* identification and analysis of SSR's from RNA-seq data of Iapar 59 and BACs sequences from the Hybrid Timor 832/2 (CaHT). For Iapar 59, 77 contigs were analyzed, 39 SSR's were found and designed 21 pairs of primers. For CaHT 77 contigs were analysed, 35 SSR's were found, and designed 29 pairs of primers. The motifs most frequently observed were di-, tri-and tetranucleotide, being $(AT)_n$ the motif most frequent between the two genotypes. This search for repetitive sequences is of great importance for future validation and increase of these markers for genetic diversity studies, mapping, and association with agronomic traits of interest.

KEY WORDS: *C. arabica*, markers SSR's, *in silico* analysis, RNA-seq, bacterial artificial chromosome.

INTRODUÇÃO

O café é uma das principais commodities agroindustriais de países em desenvolvimento e uma das mais populares bebidas não alcoólicas. Seu consumo abrange regularmente 40% da população mundial (FITTER; KAPLINSKI, 2001), sendo o Brasil e os Estados Unidos os dois maiores consumidores. O Brasil é também o maior produtor e exportador mundial de café, e a região centro-sul de minas gerais o maior estado produtor com 52,0% na produção nacional (CONAB., 2012).O cultivo no Brasil é em sua maior parte de *C. arabica*, com cerca de 70% da produção, e o restante de *C. canephora* (30%) (International Coffee Organization, http://www.ico.org; CEPLAC., 2010).
Para assegurar o poder comercial competitivo do café no país, os programas de melhoramento genético estão sempre buscando produzir cultivares mais adaptados a bióticos e abióticos que podem prejudicar a cafeicultura, assim como

investir na produção de cultivares com qualidade e sabor superior. Além disso, os mercados exigem cada vez mais esforços para o desenvolvimento de cafés de melhor de qualidade (BARUAH et al., 2003).
Estudos anteriores sobre a caracterização do germoplasma de *C. arabica*, incluindo análises usando marcadores moleculares relataram a baixa diversidade genética de *C. arabica* tanto de genótipos selvagens como cultivados (SERA et al., 2003). As justificativas para a estreita base de diversidade genética incluem a recente origem da espécie, sua reprodução preferencialmente autógama, e o limitado histórico de dispersão. Além disso, o café é uma cultura perene que exige três anos de crescimento, até a plena maturidade (MENDES E GUIMARÃES., 1998), sendo preciso pelo menos 20 anos para obter uma nova cultivar. Portanto a utilização de ferramentas como os marcadores moleculares é uma ótima opção, pois permite identificar a variabilidade genética do cafeeiro e a seleção precoce de características de interesse com menor custo e tempo. Dentre os marcadores moleculares atualmente disponíveis, Simple Sequence Repeats (SSR's) ou microssatélites, são sequências de DNA compostas por curtas repetições em tandem em um dado loco (TAUZ and RENZ., 1994) e possuem propriedades que fazem deles um dos mais informativos e versáteis marcadores de DNA usados em pesquisas genéticas de plantas (CSENCSICS; RODBECK AND HOLDEREGGER., 2010).
Visando o desenvolvimento futuro de marcadores SSR's para café, este trabalho tem como objetivo realizar uma análise *in silico* a partir de dados de transcriptoma (RNA-seq) da cultivar Iapar 59 e a partir de sequências de *C. arabica* 832/2 (Híbrido de Timor-CaHT). Nesta análise foram identificados, caracterizados e desenhados pares de primers para os motivos SSR's encontrados para posterior validação e estudos de diversidade genética, mapeamento e associação com características agronômicas de interesse, pois apesar da importância econômica de *C. arabica* poucas informações de mapeamento estão disponíveis.

MATERIAL E MÉTODOS

Foi realizado previamente o sequenciamento do transcriptoma (RNA-seq) de Iapar 59 e de sequências de BACs de CaHT832/2, utilizando a tecnologia Illumina HISEQ.
A análise *in silico* foi feita em 32.000 contigs de Iapar 59 obtido a partir de sequenciamento de transcriptoma e em 180 contigs genômicos de CaHT.
O programa utilizado para a busca de sequências repetitivas foi Gramene Ssrtool (http://www.gramene.org) desenvolvido por Cartingour e usado por Poncet et al. (2006) e Baruah et al. (2003)**.** Os parâmetros do programa foram definidos para a detecção de motivos mono-, di-, tri-, tetra-, penta- e hexanucleotídeos com número mínimo de 3 repetições. Além disso, os motivos foram filtrados para a anotação com unidade mínima de repetição descrita a seguir: 10 unidades de repetição para mono-, 5 unidades de repetição para di-, 4 unidades de repetição para tri-, e 3 unidades de repetição para tetra-, penta e hexanucleotídeos. Pelo programa Primer 3 (ROZEN and SKALESHY., 2000) como descrito por Varshney et al. (2002); Baruah et al. (2003) (http://www.fokker.wi.mit.edu/primer3/), e pelo programa Oligo Calc: Oligonucleotide Properties Calculator (http://www.basic.northwestern.edu/biotools/oligocalc.html) foi realizado o desenho e a escolha dos primers.
Os SSR's foram caracterizados de acordo com as unidades de repetição, e a frequência com que os motivos ocorreram no genoma da espécie.

RESULTADOS E DISCUSSÃO

A montagem dos dados de RNA-seq de Iapar 59 formou 32.000 contigs. Destes, 77 com tamanho entre 6.616 KB a 12.891 KB foram analisados. Foram encontradas sequências repetitivas em 39 contigs e foram desenhados 21 pares de primers. Para as sequências de HT 832/2 foram montados 180 contigs. Destes, 77 com tamanho entre 934 KB a 38.402 KB foram analisados, 35 continham sequências repetitivas e dessas, 29 pares de primers foram desenhados.
Os motivos mais abundantes obtidos da análise *in silico* dos cultivares de *C. arabica* foram di, tri e tetra, enquanto penta e hexanucleotídeos foram encontrados em números insignificantes.
Para Iapar 59 os dinucleotídeos foram mais frequentes com 57,14% do total de motivos identificados, enquanto que para CaHT os mais frequentes foram tri e tetra representando 40,74% cada (Tabela 1). Foram anotados os maiores motivos encontrados, isso por que quanto mais unidades de repetição, mais oportunidades para que o slippage ocorra durante a replicação. Portanto, loci com grandes números de repetições, são mais polimórficos (ELLEGREN., 2004). Assim, os motivos SSR's encontrados variaram de 3 a 9 unidades de repetição como representado na tabela 4.
Esses resultados estão de acordo com Morgante and Olevieri. (1993) que descreveram que repetições di e tri são amplamente distribuídas em plantas. Similarmente Aggarwal et al. (2007) relataram que em geral os marcadores SSR's desenvolvidos em *Coffea* sp. são principalmente compostos por repetições di- e trinucleotídeos.
Em análise *in silico* de sequências expressas (EST's) de folhas e frutos de *C. canephora* (Poncet et al., 2004) demonstraram que motivos tri- foram mais abundantes, seguidos pelos di- e hexanucleotídeos. Poncet et al. (2006) e Pereira et al. (2011) mineraram SSR's a partir de EST's demonstrando que motivos trinucleotídeos também foram mais abundantes que dinucleotídeos como observado em nossos resultados de CaHT.

Tabela 1. Frequência dos motivos encontrados em contigs de Iapar 59 e Híbrido de Timor.

Iapar 59		
Tipo do SSR	N°	%
di-SSR's	12	57,14
tri-SSR's	9	42,86
Total	21	100
Hibrido Timor		
Tipo do SSR	N°	%
di-SSR's	5	18,52
tri-SSR's	11	40,74
tetra-SSR's	11	40,74
Total	27	100

Tanto para Iapar 59 quanto para CaHT o motivo mais abundante foi da classe $(AT)_n$ (Tabelas 2 e 3). Os resultados obtidos para ambas cultivares neste estudo reforçam o que foi descrito por Cardle et al. (2000) e La Rota et al. (2005), no qual os motivos poli AT são os mais encontrados na maioria dos genomas de plantas. Além disso, a classe $(AT)_n$ foi um dos mais frequentes dinucleotídeos relatados por Aggarwall et al. (2007).
Dos motivos encontrados em bancos de dados EST's do café, as classes $(AT)_n$, $(TC)_n$, $(GA)_n$, e $(TA)_n$ foram frequentes no trabalho de Pereira et al. (2011) além de terem demonstrado polimórficos em *C. arabica* (Tabela 2).

Tabela 2. Caracterização dos motivos SSR's encontrados para Iapar 59.

Repetição di-nucleotídeo	Número de di-SSR's	%
AT	2	25
TC	2	25
GA	2	25
TA	2	25
Total	8	100
Repetição tri-nucleotídeo	**Número de tri-SSR's**	**%**
ATC	2	100
Total	2	100

Tabela 3. Caracterização dos motivos SSR's encontrados para CaHT.

Repetição di-nucleotídeo	Número de di-SSR's	%
AT	4	100
Total	4	100
Repetição tri-nucleotídeo	**Número de tri-SSR's**	**%**
TTA	2	33,3
GTG	2	33,3
TGC	2	33,3
Total	6	100
Repetição tetra-nucleotídeo	**Número de tetra-SSR's**	**%**
TAAA	2	50
ATTT	2	50
Total	4	100

Tabela 4. Caracterização dos motivos SSR's quanto ao tamanho da unidade de repetição.

Cultivar	Motivo SSR	N° de unidade de repetição						
		3	4	5	6	7	8	9
Iapar 59	AT			1			1	
	TC				2			
	GA			1			1	
	TA			2				
	ATC		1	1				
HDT	AT				1	2		1
	TTA		2					
	GTG		1		1			
	TGC		2					
	TAAA	2						
	ATTT	2						

*quantidade de motivos SSR's encontrados

CONCLUSÕES

A análise *in silico* foi eficiente para a busca e caracterização de marcadores microssatélites para a espécie *C. arabica*. Foram desenhados 21 primers para Iapar 59 e 29 primers para Híbrido de Timor para que sejam validados aumentando assim, o número de marcadores informativos em cultivares comerciais de *C.arabica*. Assim, posteriormente os primers validados serão utilizados em estudos de diversidade genética, mapeamento e associação com características agronômicas de interesse.

REFERÊCIAS BIBLIOGRÁFICAS

AGGARWAL, R.K; HENDRE, P.S; VARSHNEY, R.K; BHAT, P.R; KRISHNAKUMAR, V AND SINGH, L. Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. Theor Appl Genet, 114:359-372, 2007.

BARUAH, A; NAIK, V; HENDRE, P.S; RAJKUMAR, R; RAJENDRAKUMAR, P; AGGARWAL, R.K. Isolation and characterization of nine microsatellite markers from *Coffea arabica* L., showing wide cross species amplifications. Molecular Ecology Notes, v.3, p.647-650, 2003.

CARDLE, L; RAMSAY, L; MILBOURNE, D; MACAULAY, M; MARSHALL, D AND WAUGH, R. Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics, 156: 847-854, 2000.

CEPLAC. Disponível em: acessado em outubro de 2010.

CSENCSICS; D.S.B RODBECK AND HOLDEREGGER, R. Cost-effective, species-specifi c microsatellite development for the endangered dwarf bulrush (Typha minima) using next-generation sequencing technology.Journal of Heredity, 101 : 789 – 793, 2010.

CONAB. Companhia Nacional de Abastecimento. Levantamentos de safra – 3º Levantamento de Café Setembro/2012 Disponível em: http://www.conab.gov.br/OlalaCMS/uploads/arquivos/12_09_06_10_10_21_boletim_cafe_setembro_2012.pdfp. 1-19, 2012.

ELLEGREN, H. Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics*, Vol.5, No.6, (June 2004), pp. 435–445, ISSN 1471-005, 2004.

FITTER, R; KAPLINSKY, R: Who gains from product rents as the coffee market becomes more differentiated? A value chain analysis. IDS Bulletin (Special Issue), 32(3):69-82, 2001.

LA ROTA, M; KANTETY, R.V; YU, J.K AND SORRELLS, M.E. Nonrandom distribution and frequencies of genomic and ESTderived microsatellite markers in rice, wheat, and barley.BMC Genomics, 6: 23, 2005.

MENDES, A.N.G AND GUIMARÃES, R.J. Genética e Melhoramento do Caffeiro. Lavras: UFLA/FAEPE, 1998.

TAUTZ, D AND RENZ, M. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Research,12: 4127-4138, 1994.

MORGANTE, M AND OLIVIERI, A.M. PCR-amplified microsatellite markers in plant genetics. Plant, J 3:175-182, 1993.
PEREIRA, G. S; PADILHA, L; VILELA, E & PINHO, R. VON. Microsatellite markers in analysis of resistance to coffee leaf miner in Arabica coffee, (1), 1650–1656, 2011.
PONCET, V; HAMON, P; MINIER, J; CARASCO, C; HAMON, S; NOIROT, M. SSR cross-amplification and variation within coffee trees (*Coffea* spp.). Genome, v.47, p.1071‑1081, 2004.
PONCET, V; RONDEAU, M; TRANCHANT, C; CAYREL, A; HAMON S, DE KOCHKO, A AND HAMON, P. SSR mining in coffee tree EST databases: potential use of EST-SSRs as markers for the Coffea genus. Molecular and Genetics Genomics, 276: 436-449, 2006.
ROZEN, S; SKALETSKY, H.J. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics methods and protocols: methods in molecular biology. Humana Press, Totowa, pp 365–386, 2000.
SERA, T; RUAS, P.M; RUAS, C.F; DINIZ, L.E.C; CARVALHO, V.P; RAMPIM, L; RUAS, E.A; SILVEIRA, S.R. Genetic polymorphism among 14 elite Coffea arabica L. cultivars using RAPD markers associated with restriction digestion. Genetics and Molecular Biology, v.26, p.59-64, 2003.
VARSHNEY, R.K; THIEL, T; STEIN, N; LANGRIDGE, P; GRANER, A. In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell Mol Biol Lett,7:537–546, 2002.
VERA, J.C; WHEAT, C.W; FESCEMYER, H.W, FRILANDER, M.J; CRAWFORD, D.L; HANSKI, I; NARDEN, J.H: Rapid transcriptome characterizarion for a non model organism using 454 pyrosequencing. Mol Ecol 17(7): 1636-1647, 2008.