

## Identificação e Caracterização de Microssatélites de *Coffea arabica* a partir de dados de sequenciamento de RNA e de BACs

**Bruna Silvestre Rodrigues da Silva<sup>1</sup>; Sandra Bellodi Cação<sup>2</sup>; Suzana Tiemi Ivamoto<sup>3</sup>; Juliana Costa Silva<sup>4</sup>; Douglas Silva Domingues<sup>5</sup>; Luiz Filipe Protasio Pereira<sup>6</sup>**

<sup>1</sup>Bolsista Consórcio Pesquisa Café, Londrina – PR, brunasilvestrerodrigues@hotmail.com

<sup>2</sup>Pesquisador, Dr, Agronomia – UEL, Londrina – PR, sandracacao@sercomtel.com.br

<sup>3</sup>Doutoranda de Genética e Biologia Molecular – UEL, Londrina – PR, suzanatiemi@yahoo.com.br

<sup>4</sup>Técnico Bolsista de Bioinformática, Iapar, Londrina – PR, jujuconstasilvaanalise@gmail.com

<sup>5</sup>Pesquisador, Dr, Instituto Agrônomo do Paraná, Londrina - PR, doug@iapar.br

<sup>6</sup>Pesquisador, Dr, Embrapa Café, Brasília – DF, filipe.pereira@embrapa.br

**RESUMO:** O café é uma das principais commodities agrícolas comercializados. A estreita base genética de *Coffea arabica* tem-se refletido em diferentes formas. Além disso, em *C. arabica*, o desenvolvimento de novas cultivares é demorado. A utilização de marcadores moleculares como ferramenta para análises da diversidade e seleção de genótipos representa uma importante ferramenta para tentar diminuir tempo e custo de seleção do melhoramento de espécies perenes como *C. arabica*. Neste trabalho foi realizada identificação e análise *in silico* de SSR a partir de dados de RNA-seq de *C. arabica*-Iapar 59 e de sequenciamento de BACs de *C. arabica* HDT832/2. Para Iapar 59 foram analisados 77 contigs, encontrados 39 SSR's e desenhados 21 primers. Para HDT832/2 foram analisados também 77 contigs, encontrados 35 SSR's e desenhados 29 primers. Os motivos mais frequentes foram di, tri e tetra, sendo AT o mais frequente entre os genótipos. Esta busca de seqüências repetitivas é de grande importância para futura validação e aumento desses marcadores para estudos de diversidade genética.

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

### 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<sup>1</sup>, 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<sup>2</sup>. 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%)<sup>3</sup>.

Para assegurar o poder comercial competitivo do café no país, os programas de melhoramento genético estão sempre buscando soluções para fatores bióticos e abióticos que podem prejudicar a cafeicultura, assim como investir na produção de cultivares com sabor superior e qualidade diferencial. Além disso, os mercados exigem cada vez mais esforços para o desenvolvimento de cafés de melhor de qualidade<sup>4</sup>.

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<sup>5</sup>. 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<sup>6</sup>, sendo preciso pelo menos 20 anos para obter uma nova cultivar. Entretanto a utilização de ferramentas como os marcadores moleculares é uma ótima opção, pois permitem acesso à variabilidade genética do cafeeiro e a escolha precoce de características de interesse com menor custo e tempo. Dentre os marcadores moleculares atualmente disponíveis, Simple Sequence Repeats (SSR's) ou microssatélites, são seqüências de DNA compostas por curtas repetições em tandem em um dado loco<sup>7</sup> e possuem propriedades que fazem deles um dos mais informativos e versáteis marcadores de DNA usados em pesquisas genéticas de plantas<sup>8</sup>.

As novas técnicas de sequenciamento (NGS - *next generation sequencing*) vêm permitindo a análise detalhada de genomas, sendo utilizadas com sucesso no sequenciamento de genomas e transcriptomas (RNA-seq)<sup>9</sup>. 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 cromossomos artificiais de bactérias (BACs) sequenciados do genótipo de *C. arabica* 832/2 (Híbrido de Timor-CaHT). Nesta análise serão 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 agrônômicas 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 seqüências de BACs de CaHT832/2, utilizando a tecnologia Illumina HISEQ.

A análise *in silico* foi feita em 32.000 contigs de Iapar 59 obtidos a partir de sequenciamento de transcriptoma e em 180 contigs genômicos de CaHT.

O programa utilizado para a busca de seqüências repetitivas foi Gramene Ssrtool (<http://www.gramene.org>) desenvolvido por Cartingour e usado por Poncet et al.<sup>10</sup> e Baruah et al.<sup>4</sup>. 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<sup>11</sup> como descrito por Varshney et al.<sup>12</sup>; Baruah et al.<sup>4</sup> (<http://www.fokker.wi.mit.edu/primer3/>), e 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 seqüências repetitivas em 39 contigs e foram desenhados 21 pares de primers. Para as seqüências de HT 832/2 foram montados 180 contigs. Destes, 77 com tamanho entre 934 KB a 38.402 KB foram analisados, 35 continham seqüências 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 tetranucleotídeos, enquanto penta e hexa 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 tetranucleotídeos 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<sup>13</sup>. 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.<sup>14</sup> que descreveram que repetições di e tri são amplamente distribuídas em plantas. Similarmente Aggarwal et al.<sup>15</sup> relataram que em geral os marcadores SSR's desenvolvidos em *Coffea* sp. são principalmente compostos por repetições dinucleotídeos e trinucleotídeos.

Em análise *in silico* de seqüências expressas (EST's) de folhas e frutos de *C. canephora*<sup>16</sup> demonstraram que motivos tri- foram mais abundantes, seguidos pelos di- e hexanucleotídeos. Poncet et al.<sup>11</sup> e Pereira et al.<sup>17</sup> que mineraram SSR's à partir de EST's discutiram em seus

resultados 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
Híbrido 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 e CaHT o motivo mais abundante foi da classe (AT)<sub>n</sub> como representado nas tabelas 2 e 3. Os resultados obtidos para ambas cultivares neste estudo reforçam o que foi descrito por La Rota et al.<sup>18</sup> e Cardle et al.<sup>19</sup>, no qual os motivos poli AT são os mais encontrados na maioria dos genomas de plantas. Além disso, a classe (AT)<sub>n</sub> foi um dos mais frequentes dinucleotídeos relatados por Aggarwall et al.<sup>15</sup>.

Além disso, dos motivos encontrados em bancos de dados EST's do café, as classes (AT)<sub>n</sub>, (TC)<sub>n</sub>, (GA)<sub>n</sub>, e (TA)<sub>n</sub> foram frequentes no trabalho de Pereira et al.<sup>16</sup> além de terem demonstrado polimorfismos claros em *C. arabica* (Tabela 2).

Tabela 2. Caracterização dos motivos 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 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 microsatélites para a espécie *C. arabica*. Foram desenhados um total de 21 primers para Iapar 59 e 29 primers para Híbrido de Timor, para que sejam validados na espécie aumentando assim, o número de primers informativos em cultivares comerciais de *C. arabica*. Assim, posteriormente os primers validados serão utilizados em estudos de diversidade genética, mapeamento e associação com características agrônômicas de interesse.

## REFERÊNCIAS BIBLIOGRÁFICAS

- (1) 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.
- (2) 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.pdf](http://www.conab.gov.br/OlalaCMS/uploads/arquivos/12_09_06_10_10_21_boletim_cafe_setembro_2012.pdf). 1-19, 2012.
- (3) VIDAL, R.O; MONDEGO, J.M.C; POT, D; AMBRÓSIO, A.B; ANDRADE, A.C; PEREIRA, L.F.P; COLOMBO, C.A; VIEIRA, L.G.E; CARAZZOLLE, M.F; PEREIRA, G.A.G. A High Throughput data mining of single nucleotide polymorphisms in *Coffea* species expressed sequence tags suggests differential homeologous gene expression in the allotetraploid *Coffea arabica*. **Plant Physiology**, 154:1053-1066, 2010.
- (4) 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.
- (5) 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.
- (6) MENDES, A.N.G AND GUIMARÃES, R.J. Genética e Melhoramento do Caffeiro. Lavras: UFLA/FAEPE, 1998.
- (7) TAUTZ, D AND RENZ, M. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. **Nucleic Acids Research**, 12: 4127-4138, 1994.
- (8) CSENSICS; D.S.B RODBECK AND HOLDEREGGER, R. Cost-effective, species-specific microsatellite development for the endangered dwarf bulrush (*Typha minima*) using next-generation sequencing technology. **Journal of Heredity**, 101 : 789 – 793, 2010.
- (9) VERA, J.C; WHEAT, C.W; FESCEMYER, H.W, FRILANDER, M.J; CRAWFORD, D.L; HANSKI, I; NARDEN, J.H: **Rapid transcriptome characterization for a non model organism using 454 pyrosequencing.** **Mol Ecol** 17(7): 1636-1647, 2008.
- (10) 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.

- (11) 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.
- (12) 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.
- (13) ELLEGREN, H. Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics*, Vol.5, No.6, (June 2004), pp. 435–445, ISSN 1471-005, 2004.
- (14) MORGANTE, M AND OLIVIERI, A.M. PCR-amplified microsatellite markers in plant genetics. **Plant, J** 3:175-182, 1993.
- (15) 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.
- (16) 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.
- (17) 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.
- (18) 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.
- (19) 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.