# DIVERSIDADE NUCLEOTÍDICA DE GENES ENVOLVIDOS NA BIOSSÍNTESE DE ÁCIDOS CLOROGÊNICOS DE CAFEEIROS

Suzana Tiemi Ivamoto[1], David Pot[2], Sergio Dias Lannes[3], Douglas Silva Domingues[4], Luiz Gonzaga Esteves Vieira[5], Luiz Filipe Protasio Pereira[6]

(Recebido: 20 de outubro de 2011; aceito: 20 de março de 2012)

**RESUMO**: Os ácidos clorogênicos (CGAs) são compostos químicos importantes de *Coffea* spp. para a qualidade da bebida, pois eles interferem na adstringência e podem alterar o aroma e sabor da bebida. Aproximadamente 310.000 ESTs de *Coffea* estão disponíveis e possibilitam o acesso à variabilidade nucleotídica da planta e o desenvolvimento de marcadores moleculares ligados à qualidade da bebida para as principais enzimas da via de biossíntese dos CGAs: PAL, C4H, 4CL, CQT e C3'H. Neste trabalho foram detectados polimorfismos dos tipos SNP, INDEL ou SSR dentro das sequências nucleotídicas disponíveis no Projeto Genoma Café e no NCBI. As sequências de ESTs de CGAs foram clusterizadas pelo programa Codon Code Aligner, assim como a detecção de polimorfismos e validação dos mesmos (qualidade de cromatograma). Foram identificadas seis isoformas para PAL, uma para C4H, seis para 4CL, duas para CQT e duas para C3'H. Os contigs formados apresentaram um total de 248 polimorfismos (236 SNPs e 12 INDELs), sendo 201 na região codante (127 não sinônimos e 74 sinônimos). A frequência dos polimorfismos foi maior nas regiões UTRs (1pol/54pb), em relação à codante (1pol/81pb). A análise das sequências de *C. arabica* permitiu a identificação de 2 subgrupos diferentes de sequências, referentes aos seus genomas ancestrais (*C. canephora* e *C. eugenioides*). Foi observada a presença de 67,4% dos polimorfismos entre os grupos ancestrais e 32,6% dentro dos grupos em *C. arabica.* Esses resultados vêm permitindo definir genes tanto para estudos de expressão de homeólogos de CGAs como para o desenvolvimento de marcadores moleculares para o mapeamento genético.

**Termos para indexação:** Polimorfismos; marcadores moleculares; SNPs, SSRs, ESTs, CGAs.

# NUCLEOTIDE DIVERSITY OF GENES RELATED TO CHLOROGENIC ACID BIOSYNTHESIS OF *COFFEA*

***ABSTRACT:*** *Chlorogenic acids (CGAs) are important chemical compounds of Coffea spp. related to beverage quality as they affect its astringency and can change its aroma and flavor. About 310,000 Coffea Expressed Sequence Tags (ESTs) are available and provide access to the nucleotide variability of the plant and to the development of molecular markers linked to beverage quality for the main enzymes involved in biosynthesis of the CGAs: PAL, C4H, 4CL, CQT and C3'H. In this study we identified SNP, INDELS and SSR polymorphisms within the nucleotide sequences available from the Brazilian Coffee Genome database and from the NCBI. The EST sequences for CGAs were trimmed and clustered by the program Codon Code Aligner, and polymorphisms and their validation detected (chromatogram quality). We identified six isoforms for PAL, one for C4H, six for 4CL, two for CQT and two for C3'H. The contigs formed exhibited a total of 248 polymorphisms (236 SNPs and 12 INDELs), with 201 in the coding region (127 non-synonymous and 74 synonymous). The frequency of polymorphisms was greater in the UTR regions (1pol/54pb) in relation to the coding region (1pol/81pb). The analysis of C. arabica sequences allowed identification of two different subgroups of sequences, related to their ancestral genomes (C. canephora and C. eugenioides). The presence of 67,4% of the polymorphisms between the ancestral groups and 32,6% within the groups were observed em C. arabica . The characterization of nucleotide diversity on those genes is essential for further studies on differential expression of their homeologs, as well as the use of CGAs as molecular markers related to genetic mapping.*

***Index terms:*** *Polymorphisms, molecular markers, SNPs, SSRs; ESTs, CGAs.*

## 1 INTRODUÇÃO

O cafeeiro pertence à família Rubiaceae e ao gênero *Coffea* da qual se conhecem aproximadamente 124 espécies. Dentre essas espécies duas possuem uma mair importância econômica, a *Coffea arabica* L. e *Coffea canephora* Pierre ex A. Froehner que representam 70% e 30% do mercado total de café, respectivamente (CECAFE, 2011). A espécie *C. arabica* é originária do oeste da África (Etiópia) e é uma planta alotetraplóide (2n=4x=44), as demais espécies

[1]Universidade Estadual de Londrina/UEL - Departamento de Biologia Geral - Rodovia Celso Garcia Cid - Km 380 - 86051-980 - Londrina-PR - suzanatiemi@yahoo.com.br
[2]Centre de Coopération Internationale en Recherche Agronomique pour le Development/CIRAD - TA 80/03 Avenue d'Agropolis, 34398 - Montpellier- França - Cedex 5 - david.pot@cirad.fr
[3]Empresa de Pesquisa Agropecuária de Minas Gerais/EPAGRI - Rodovia Admar Gonzaga - Km 1347 - Cx. P. 502 - 88034-901 Florianópolis - SC - sergiolannes@epagri.sc.gov.br
[4]Instituto Agronômico do Paraná/IAPAR - Rodovia Celso Garcia Cid - Km 375 - Cx P. 481 - 86001-970 Londrina - PR - doug@iapar.br
[5]Instituto Agronômico do Paraná/IAPAR - Rodovia Celso Garcia Cid - Km 375 - Cx P. 481 - 86001-970 - Londrina - PR lvieira@iapar.br
[6]Empresa Brasileira de Pesquisa Agropecuária/EMBRAPA Café - Rodovia Celso Garcia Cid Km 375 - Cx P. 481 - 86001-970 Londrina, PR - filipe.pereira@embrapa.br

são diplóides (2n=2x=22) e alógamas. A origem provável de *C. arabica* foi uma hibridização de *C. eugenioides* S. Moore e *C. canephora* (DAVIS et al., 2007; LASHERMES et al., 1999).

O Brasil é o maior produtor e exportador mundial de café, sendo responsável por 30% do mercado internacional (CECAFE, 2011). A qualidade da bebida de café é o principal fator de agregação de valor ao produto, pois confere melhores preços no mercado e, portanto maior competitividade. A composição química do café é um dos fatores que determinam a qualidade da bebida. Seu sabor e aroma são resultantes da presença combinada de vários constituintes químicos voláteis e não voláteis, entre eles os açúcares, cafeína, trigonelina, lipídeos e os CGAs (FARAH et al., 2006). Esses últimos são responsáveis pela adstringência e interferem no seu sabor (KIM; BEPPU; KATAOKA, 2009).

Os CGAs são compostos de metabolismos secundários e derivam da biossíntese da fenilalanina pela via dos fenilpropanóides (CAMPA et al., 2004). Existem duas vias de síntese dos CGAs, primeiro a enzima fenilalanina amonialiase (PAL) catalisa a fenilalanina em ácido cinâmico, a cinamato 4 hidroxilase (C4H) catalisa a segunda hidroxilação do ácido cinâmico para ácido coumárico, esse é catalizado pela enzima 4 couramato CoA ligase (4CL) em p-coumaril CoA. A partir desse ponto podem ser feitas duas rotas diferentes. Na primeira o ácido p-coumárico pode ser catalisado em ácido cafeico pela 4-couramato 3-hidroxilase (C3`H) que, em seguida, é catalisado pela enzima 4CL em Cafeoil CoA, no qual é adicionada a molécula de ácido quínico pela enzima hidroxicinamoil transferase quinato (CQT), originando as moléculas de CGAs. Na segunda, o ácido p-coumárico é adicionado à molécula de ácido quínico pela enzima CQT que origina o ácido p-coumaroil quínico, o qual é catalisado pela enzima C3'H em CGAs (RUPASINGUE, 2008) (Figura 1).

O estudo das enzimas envolvidas nessa via de biossíntese é importante, pois os CGAs são precursores da síntese de lignina que está envolvida na proteção das células da planta contra estresses bióticos e abióticos (resistência a patógenos e redução da permeabilidade da parede celular em relação à água), além de serem fundamentais para a estruturação rígida das plantas (KOCHKO et al., 2003; MCCARTHY et al., 2007). O interesse do uso de CGAs na saúde humana está aumentando devido às suas propriedades antioxidantes, efeito antagonista a opióides e a sua capacidade de transporte de glicose no combate ao diabetes (BAKURADZE et al., 2011).

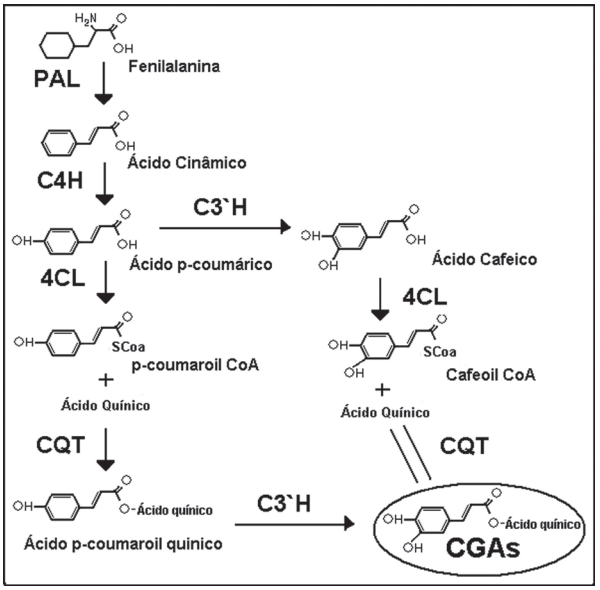


**FIGURA 1 –** Via de Biossíntese dos CGAs e as cinco enzimas envolvidas.

Com o desenvolvimento da biologia molecular, surgiram novas técnicas para auxiliar o melhoramento genético. É o caso dos marcadores moleculares que podem indicar com precisão a variabilidade genética dos indivíduos (BORÉM, 2009). Esses possuem inúmeras vantagens em relação aos marcadores morfológicos, pois independem das condições ambientais e estádio fisiológico da planta, além de permitirem uma identificação precoce de genótipos com características de interesse (EVANS; CARDON, 2004). O mapeamento de genes relacionados aos CGAs tem sido realizado em várias espécies de plantas como a maçã (CHAGNE et al., 2012) e alcachofra (MENIN et al., 2010).

Dentre os vários tipos de marcadores moleculares disponíveis, os mais interessantes para o estudo de mapeamento genético são os microssatélites (SSRs) e os polimorfismos de base única (SNPs). Esses possuem uma maior frequência dentro do genoma dos organismos e permitem a identificação e mapeamento de genes que controlam características de interesse agronômico, fator interessante para reduzir o tempo e o custo do melhoramento clássico de plantas (BORÉM, 2009).

A utilização de dados de projetos de transcriptoma de café, como ESTs (sequências de DNA expressas), vem permitindo estudos

visando entender melhor o funcionamento genético da planta (MONDEGO et al., 2011; VIDAL et al., 2010). No *Genbank* (NCBI) além das mais de 260.000 sequências do Genoma Café, estão disponíveis aproximadamente 47 mil sequências ESTs de *C. canephora* (LIN et al., 2005) e 10 mil provenientes do IRD (Institute de Recherche pour le Developpment) (PONCET et al., 2006).

Este trabalho visa a busca e seleção *in silico* de sequências das enzimas envolvidas na biossíntese dos CGAs, objetivando-se identificar o número de isoformas para cada enzima, avaliar o nível de polimorfismos disponível dentro dos ESTs, analisar a origem desses polimorfismos (inter específico ou intra específico), calcular a frequência dos mesmos de acordo com as regiões (UTRs ou codantes) e iniciar um trabalho de mapeamento das isoformas identificadas dentro de populações de *Coffea* spp. baseado em análises prévias utilizando ferramentas de bioinformática.

## 2 MATERIAL E MÉTODOS

As buscas das sequências de interesse foram feitas na plataforma do Genoma Café (http://www.lge.ibi.unicamp.br/cafe) através das ferramentas *BlastX* e no *Genbank*/NCBI (http://www.ncbi.nlm.nih.gov/) por *tBlastN* com a utilização de sequências proteicas codificantes para as enzimas envolvidas na biossíntese de CGAs previamente identificadas no genoma de *Arabidopsis thaliana* (L.) Heynh. Após comparação, foram selecionadas somente sequências com *e-value* menor que $e^{-10}$. As sequências encontradas foram inseridas no Programa *Codon Code Aligner* (versão 1.6.3) para análise e formação dos *contigs*. O mesmo programa fez a limpeza das sequências o e análise da qualidade dos cromatogramas através das ferramentas denominadas *call base*, *clip ends*, *trim vector* e *find heterozygous* (Phred e Phrap) para excluir regiões de baixa qualidade e descartar as sequências que correspondem a vetores. Os parâmetros para o alinhamento das sequências foram a presença de percentual mínimo de identidade de 90% e homologia mínima de 20 pares de base. Após a formação dos *contigs*, a sequência consenso dos mesmos foi analisada através da ferramenta *BlastX* no *Genbank*/NCBI para validar a codificação da proteína de interesse. Após confirmação, o quadro de leitura da proteína foi determinado pela análise realizada na plataforma online EXPASY Translate Tools (2012).

A identificação do número de isoformas para cada enzima foi definido através da comparação das sequências protéicas dos seus respectivos *contigs* com o uso da ferramenta bl2seq (blast two sequence/*Genbank*/NCBI), onde foram consideradas isoformas diferentes quando a porcentagem de homologia e identidade apresentavam valores menores que 90%.

A detecção dos polimorfismos foi feita baseada na qualidade dos cromatogramas e somente quando eram detectados em pelo menos duas sequências (mínimo de quatro sequências por contig). Foram desenhados oligonucleotídeos específicos para analisar preferencialmente as regiões com maior frequência polimórfica observada *in silico*.

## 3 RESULTADOS E DISCUSSÃO

A busca *in silico* de polimorfismos para as enzimas PAL, C4H, 4CL, CQT e C3'H e formação das suas isoformas, selecionou um total de 426 sequências ESTs do banco de dados do Projeto Genoma Café. O número total de ESTs para cada uma das cinco enzimas de CGAs, assim como a quantidade de contigs formados e número de isoformas encontradas estão descritas na Tabela 1.

Análises de BlastX dos 16 contigs formadas no *Genbank* (NCBI) resultaram em similaridade com algumas sequências previamente anotadas para codificação das proteínas relacionadas ao metabolismo de CGAs (Tabela 2). A validação com os dados do *Genbank* é importante visando confirmar a montagem e tamanho dos contigs, assim com sua classificação dentro dos CGAs. Todas as isoformas encontradas nesse trabalho foram depositadas no *Genbank* (NCBI). É interessante observar que o número de ESTs das duas primeiras enzimas da via de fenilpropanóides, PAL, C4H e 4CL, foi três vezes maior que o das duas ultimas enzimas envolvidas na formação dos CGAs. Essas enzimas respondem não somente para produção de CGAs, mas também para a produção de uma série de outros compostos como, por exemplo, flavonóides, ácido salicílico e estilbenos.

Os números totais de sequências ESTs selecionadas pelas buscas *in silico* para cada contig estão representados na Tabela 3. Essas sequências foram separadas de acordo com as suas respectivas espécies de origem pelo fato de influenciarem as porcentagens de frequência dos polimorfismos.

**TABELA 1 –** Seleção de ESTs e caracterização dos contigs e singletons encontrados in silico.

| ENZIMA | ESTs | Contigs* | Singletons | Isoformas | Proteína | |
|---|---|---|---|---|---|---|
| | | | | | Completa | Incompleta |
| PAL | 133 | 4 | 2 | 5 | 2 | 2 |
| C4H | 89 | 1 | 0 | 1 | 1 | 0 |
| 4CL | 114 | 6 | 0 | 6 | 3 | 3 |
| CQT | 47 | 3 | 0 | 2 | 2 | 1 |
| C3'H | 46 | 2 | 0 | 2 | 2 | 0 |

* Contigs formados com, no mínimo, 4 sequências de ESTs.

**TABELA 2 –** Resultado do BlastX dos 16 contigs comparados contra o GenBank (NCBI).

| Enzimas | Tamanho dos contigs | Organismos | Número de acesso (NCBI) | E-value |
|---|---|---|---|---|
| PAL_C1 | 2673 pb | *Coffea canephora* | AAN32866.1 | 0.0 |
| PAL_C2 | 1056 pb | *Coffea arabica* | AEL21617.1 | 1 e-130 |
| PAL_C3 | 2253 pb | *Coffea canephora* | AEO94540.1 | 0.0 |
| PAL_C4 | 2489 pb | *Coffea canephora* | AEO94541.1 | 0.0 |
| C4H_C4 | 3349 pb | *Catharanthus roseus* L. | CAA83552 | 0.0 |
| 4CL_C2 | 2389 pb | *Coffea arabica* | CAJ41420.1 | 0.0 |
| 4CL_C3 | 2065 pb | *Rubus idaeus* L. | AAF91309.1 | 0.0 |
| 4CL_C4 | 2109 pb | *Nicotiana tabacum* L. | AAB18637.1 | 0.0 |
| 4CL_C5 | 1543 pb | *Arabidopsis thaliana* | AAP03021.1 | 0.0 |
| 4CL_C6 | 1083 pb | *Nicotiana sylvestris* Speg. & S. Comes | AAO25512.1 | 1 e-137 |
| 4CL_C7 | 1307 pb | *Populus trichocarpa* Torr. S. & Gray | EEE96927.1 | 3 e-172 |
| CQT_C1 | 1097 pb | *Coffea canephora* | ABO77957.1 | 2 e-62 |
| CQT_C4 | 2169 pb | *Coffea arabica* | CAT00081.1 | 0.0 |
| CQT_C14 | 1650 pb | *Coffea canephora* | ABO47805.1 | 0.0 |
| C3`H_C25 | 1939 pb | *Coffea canephora* | ABO77958.1 | 0.0 |
| C3`H_C34 | 1698 pb | *Coffea canephora* | ABO83677 | 0.0 |

De acordo com os dados apresentados na tabela 3, é possível perceber que, em quase todos os contigs há uma predominância de sequências de *C. arabica*, com exceção do Contig 3 da PAL que apresentou um equilíbrio, indicando que essa isoforma deve possuir uma maior expressão em *C. canephora*. Esse resultado também é devido à presença majoritária de bibliotecas de *C. arabica* no Projeto Genoma Café (VIEIRA et al., 2006).

Nos 16 contigs foram identificados um total de 248 polimorfismos, sendo 47 dentro das regiões UTRs e 201 dentro das regiões codantes. Inseridos nessa última 63% dos polimorfismos (127) correspondem ao tipo sinônimo (S) e 37% deles (74) ao tipo não sinônimo (NS) (Tabela 4). Uma porcentagem semelhante foi verificada para trigo por RAVEL et al. (2006) que encontrou aproximadamente 33% de S e 67% de NS, demonstrando que a maioria das mutações mais comuns não alteram o tipo do aminoácido formado.

Os polimorfismos foram encontrados tanto em regiões não codificadoras de proteínas (5`UTR e 3`UTR), como em regiões codificadoras. Os resultados evidenciam a maior frequência de SNPs (95%) em relação aos indels (5%), assim como a de polimorfismos sinônimos (63%) em relação aos não sinônimos (37%), como foi observado por SCHMID et al. (2003).

A partir da detecção dos polimorfismos calcularam-se as frequências da presença de polimorfismos nas 3 diferentes regiões das sequências (5`UTR, região codificadora da proteína e 3`UTR) a cada 100 pares de bases, de acordo com o seu respectivo contig. Foi encontrada frequência de 1 polimorfismo a cada 50 pares de base nas regiões UTRs e 1 polimorfismo a cada 81 pares de base, nas regiões codificadoras de proteína.

**TABELA 3 –** Número total de seqüências e distribuição nas espécies de *Coffea.*

| Enzima | Sª Total | Sª (Cc) | Sª (Ca) | Sª (Cr) |
|---|---|---|---|---|
| PAL_C1 | 34 | 6 | 27 | 1 |
| PAL_C2 | 7 | 4 | 3 | 0 |
| PAL_C3 | 52 | 27 | 25 | 0 |
| PAL_C4 | 38 | 6 | 31 | 1 |
| C4H_C4 | 89 | 21 | 67 | 1 |
| 4CL_C2 | 60 | 18 | 41 | 1 |
| 4CL_C3 | 19 | 6 | 13 | 0 |
| 4CL_C4 | 20 | 5 | 15 | 0 |
| 4CL_C5 | 6 | 1 | 5 | 0 |
| 4CL_C6 | 2 | 0 | 2 | 0 |
| 4CL_C7 | 7 | 2 | 4 | 0 |
| CQT_C1 | 2 | 0 | 2 | 0 |
| CQT_C4 | 37 | 12 | 25 | 0 |
| CQT_C14 | 8 | 4 | 4 | 0 |
| C3`H_C25 | 42 | 17 | 25 | 0 |
| C3`H_C34 | 4 | 0 | 4 | 0 |

Sª = Número Total de Seqüências; Cc = *Coffea canephora*; Ca = *Coffea arabica*; Cr= *Coffea racemosa* Ruiz & Pav.

A pesquisa realizada *in silico* para se conhecer a frequência dos polimorfismos (Tabela 5) de acordo com a sua localização é importante, pois pode-se inferir a região que possui a maior probabilidade de encontro dos mesmos nas análises *in vivo*. Com essa infomação, é possível realizar o desenho dos oligonucleotídeos específicos, priorizando as regiões que apresentem uma frequência maior de polimorfismos para a detecção dos mesmos. Segundo LAI et al. (2008), em análises de ESTs de girassóis, a porcentagem de sucesso de encontrar um polimorfismo *in vivo* foi 72%, quando análises prévias *in silico* indicaram as melhores regiões para desenhos de oligonucleotídeos, contra 37% de sucesso quando eles eram desenhados ao acaso.

Os dados de frequência dos polimorfismos evidenciaram que, na maioria das vezes, esssa encontra-se nas regiões não codificadoras de proteína (3`UTR e 5`UTR), assim como foi relatado por MONDEGO et al. (2011).

A grande frequência de SNPs encontrados em todas as enzimas estudadas confirma os dados relatados por LIJAVETZKY et al. (2007) que diz que os SNPs são os polimorfismos mais frequentes no genoma de uva. A média de SNPs totais encontrados foi de 1.4 polimorfismos a cada 100 pares de base, uma quantidade relativamente alta se comparada aos resultados obtidos por VIDAL et al. (2010) que observou 0.39 polimorfismos, porém isso provavelmente deve-se ao fato de que a região 5'UTR da C3'H_C25 elevou a média observada. Outro estudo que comprova a maior frequência dos SNPs em relação aos microssatélites (SSRs) foi o realizado por PONCET et al. (2006) que encontrou um SSR a cada 7730pb em ESTs de *C. canephora*, valor quase 5.5 vezes menor do que os SNPs detectados por este trabalho.

**TABELA 4 –** Quantidade e tipos de polimorfismos encontrados.

| Enzima | Número de Polimorfismos Totais | SNPs | Indels | Região 5'UTR | Região Codificadora | | Região 3'UTR |
|---|---|---|---|---|---|---|---|
| | | | | | Sinônimos | Não Sinônimos | |
| PAL_C1 | 23 | 23 | 0 | 0 | 18 | 5 | 0 |
| PAL_C2 | 2 | 2 | 0 | 1 | 1 | 0 | 0 |
| PAL_C3 | 43 | 42 | 1 | 0 | 22 | 17 | 4 |
| PAL_C4 | 32 | 32 | 0 | 0 | 20 | 8 | 4 |
| C4H_C4 | 34 | 30 | 4 | 0 | 17 | 11 | 6 |
| 4CL_C2 | 22 | 20 | 2 | 0 | 6 | 5 | 11 |
| 4CL_C3 | 4 | 3 | 1 | 0 | 2 | 1 | 1 |
| 4CL_C4 | 25 | 25 | 0 | 3 | 15 | 7 | 0 |
| 4CL_C6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4CL_C7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CQT_C4 | 24 | 24 | 0 | 1 | 15 | 6 | 2 |
| CQT_C14 | 4 | 4 | 0 | 0 | 2 | 2 | 0 |
| C3`H_C25 | 35 | 31 | 4 | 13 | 9 | 12 | 1 |

**TABELA 5 –** Frequência dos Polimorfismos segundo suas localizações.

| Gene | Frequência 5'UTR/100 pb | Frequência Coding/100 pb | Frequência 3'UTR/100 pb | Frequência Média |
|---|---|---|---|---|
| PAL_C1 | NA | 1,07 | 0 | 1pol/99 bp |
| PAL_C2 | 0,50 | 0,25 | NA | 1pol/292 bp |
| PAL_C3 | NA | 2,22 | 1,65 | 1pol/46 bp |
| PAL_C4 | 0 | 1,35 | 3,30 | 1pol/78 bp |
| C4H_C4 | NA | 1,80 | 3,52 | 1pol/54 bp |
| 4CL_C2 | NA | 0,69 | 2,82 | 1pol/93 bp |
| 4CL_C3 | NA | 0,22 | 0,86 | 1pol/429 bp |
| 4CL_C4 | 3,16 | 1,35 | 0 | 1pol/71bp |
| 4CL_C6 | 0 | NA | NA | 0 |
| 4CL_C7 | NA | 0 | 0 | 0 |
| CQT_C4 | 1,03 | 1,62 | 0,60 | 1pol/78 bp |
| CQT_C14 | NA | 0,41 | 0 | 1pol/52 bp |
| C3`H_C25 | 12,38 | 1,37 | 0,51 | 1pol/52 bp |

NA= regiões com menos de quatro sequências.

De acordo com os resultados obtidos na detecção de polimorfismos e identificação dos mesmos, pode-se inferir a existência de diferenças alélicas interespécies (*C. arabica* x *C. canephora*, *C. arabica* x *C. racemosa* e *C. canephora* x *C. racemosa*) e intraespécies (*C. arabica* x *C. arabica*, *C. canephora x C. canephora* e *C. racemosa* x *C. racemosa*). Os tipos de polimorfismos SNPs, detectados na análise *in silico* para as cinco enzimas estão contidas na Tabela 6.

A análise dos polimorfismos inter (E) e intra grupos (I) resultou no encontro de 71% de polimorfismos dentro de *C. arabica*, 18% dentro de *C. canephora*, 3.9% entre *C. arabica* e *C. canephora*, 1.4% entre *C. arabica* e *C. eugenioides* e 5.6% entre *C. canephora* e *C. eugenioides*. Esses dados corroboram com os encontrados por VIDAL et al. (2010), cujo estudo encontrou 81% dos polimorfismos dentro de *C. arabica*, 17% dentro de *C. canephora*, 4% entre *C. arabica* e *C. canephora*, 2.5% entre *C. arabica* e *C. eugenioides*.

Também foram classificados os polimorfismos entre as espécies (31 polimorfismos) e dentro delas (253 polimorfismos). Dentro das espécies foi observado um maior número de polimorfismos em *C. arabica* (71%), seguido de *C. canephora* (18,3%). A maioria dos SNPs encontrados localizaram-se dentro da espécie *C. arabica*, isso devido ao maior número de sequências dessa espécie, dentro da maioria dos contigs.

**TABELA 6** – Tipos de polimorfismos encontrados *in silico* para as enzimas dos CGAs.

| Gene | Contig | Intra Ca | Intra Cc | Intra Cr | Inter Ca e Cc | Inter Ca≠Cr | Inter Cc≠Cr |
|---|---|---|---|---|---|---|---|
| PAL | 1 | 17 | 5 | NA* | 2 | 0 | 2 |
| PAL | 2 | 1 | 2 | NA | 0 | NA | NA |
| PAL | 3 | 32 | 20 | NA | 0 | NA | NA |
| PAL | 4 | 28 | 0 | NA* | 0 | 4 | 6 |
| C4H | 1 | 26 | 4 | NA* | 4 | 0 | 7 |
| 4CL | 1 | 18 | 8 | NA* | 1 | 0 | 1 |
| 4CL | 2 | 0 | 1 | NA | 3 | NA | NA |
| 4CL | 3 | 23 | 3 | NA | 0 | NA | NA |
| 4CL | 4 | 0 | 0 | NA | 0 | NA | NA |
| 4CL | 6 | 0 | 0 | NA | 0 | NA | NA |
| CQT | 2 | 22 | 2 | NA | 0 | NA | NA |
| CQT | 3 | 3 | 0 | NA | 1 | NA | NA |
| C3`H | 1 | 31 | 7 | NA | 0 | NA | NA |
| TOTAL | | 201 | 52 | 0 | 11 | 4 | 16 |

Ca = *C. arabica*; Cc= *C. canepora*; Cr = *C. racemosa;* NA= não havia sequências de *C. racemosa.* NA*= Havia apenas 1 sequência de *C. racemosa.*

Dentro da espécie de *C. arabica* foi possível separar as sequências em 2 grupos genômicos diferentes (subgenomas), um deles assemelha-se grupo de *C. canephora* e o outro provavelmente seja semelhante a *C. eugenioides*. Isso se deve, possivelmente, ao fato de que *C. arabica* é um híbrido originado entre o cruzamento de *C. canephora* e *C. eugenioides* (Figura 2).

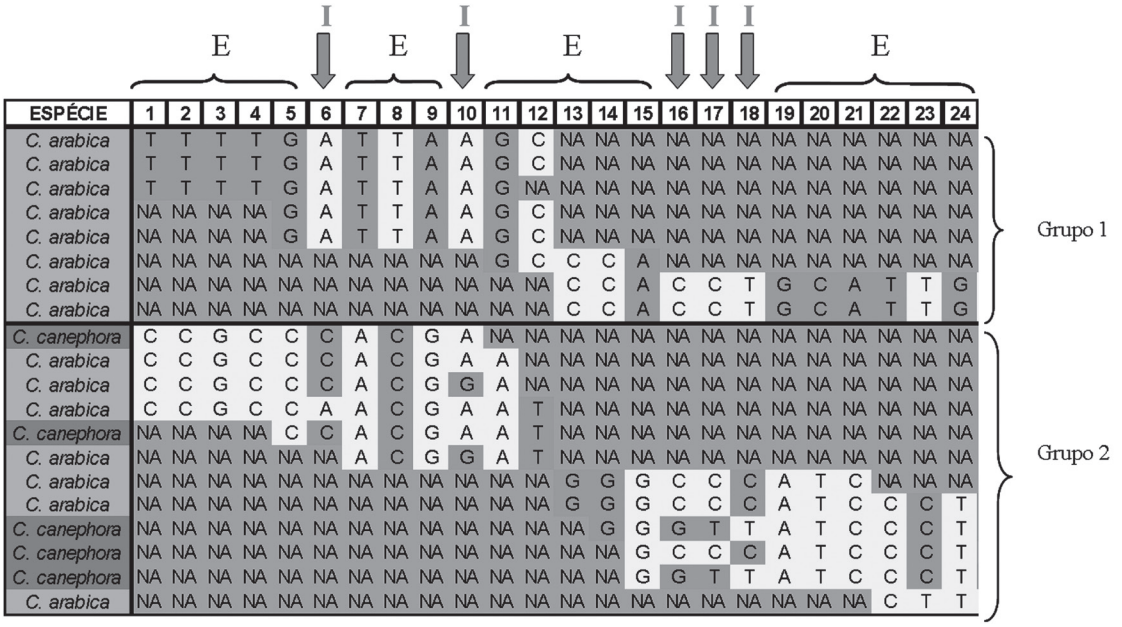


**FIGURA 2** – Exemplo de Identificação e Caracterização de Polimorfismos Intra-espécies e Inter-espécies (Contig4 da 4CL). E: polimorfismos entre grupos de *Coffea arabica*. I: polimorfismos dentro dos grupos de *C. arabica*, indicados com seta azul. NA: ausência de alelos. A: adenina. C: citosina. G: guanina. T: timina. Grupo 1: grupo de *C. arabica* semelhante à *Coffea eugenioides*. Grupo 2: grupo de *C. arabica* semelhante à *Coffea canephora* e para se observar melhor essa homologia, inseriu-se no grupo 2 algumas sequências de *C. canephora*.

A formação de 2 grupos distintos de sequências dentro de *C. arabica*, reforçam os trabalhos sobre a origem alopoliplóide da espécie a partir da hibiridização das duas espécies diplóides *C. canephora* e *C. eugenioides* (DAVIS et al., 2007; LASHERMES et al., 1999).

Analisando apenas os polimorfismos dentro dos subgenomas de *C. arabica*, foi possível concluir que a maioria deles situava-se entre os dois grupos (136 polimorfismos) e a minoria dentro de um dos dois grupos formados (68 polimorfismos). Essa separação por SNPs é importante, pois permite estudos de caracterização da expressão de genes homeólogos em *Coffea arabica*, identificando a expressão diferencial de genes de um subgenoma com relação ao outro (MARRACCINI et al., 2011; VIDAL et al., 2010). Entretanto visando a identificação de genótipos diferentes e estudos de mapeamento, somente os SNPs intraespecíficos podem ser utilizados.

Os resultados gerados com a utilização de ESTs disponíveis nos bancos de dados públicos é um trabalho de base que auxilia e aumenta as probabilidades de se desenvolverem marcadores moleculares efetivos para características de interesse agronômico, pois possibilita inferir a diversidade nucleotídica esperada *in vivo* de indivíduos de uma população, acelerando o processo de busca de marcadores polimórficos. Esse fator é de fundamental importância para futuros trabalhos de mapeamento e de seleção assistida por marcadores de espécies perenes economicamente importantes, diminuindo custos e tempo de melhoramento clássico da planta (COGAN et al., 2007).

Baseado nessas análises *in silico*, algumas populações de *Coffea* sp. estão sendo genotipadas com os *primers* desenvolvidos por este trabalho. Sendo que um deles já apresentou resultados positivos e foi mapeado em *Coffea canephora* (LEROY et al., 2011). Para a população de mapeamento de *Coffea arabica* as análises estão em desenvolvimento.

## 4 CONCLUSÕES

As análises *in silico* são eficientes para identificar genes das principais vias metabólicas dos CGAs de cafeeiro. Foram identificados 16 contigs relacionados a cinco dos principais genes da biossíntese de CGAs.Esses genes apresentam uma alta frequência de polimorfismos do tipo SNPs (95%), com sua maior distribuição nas regiões UTRs (1pol/50pb) em relação às regiões codificadoras de proteína (1pol/81pb). A análise *in silico* também identificou a presença

de dois grupos distintos de sequências dentro do grupo de *C. arabica*, de acordo com os seus genomas ancestrais, permitindo realizar estudos de expressão de genes homeólogos. Os genes e polimorfismos encontrados vêm sendo utilizados em trabalhos de genotipagem de populações de *Coffea* sp., para mapeamento físico e genético dos CGAs.

## 5 AGRADECIMENTOS

Ao Consórcio Pesquisa Café e FINEP/GENOCAFÉ, pelo apoio financeiro. S.T.I. foi bolsista de Mestrado CNPq; L.F.P.P. e L.G.E.V. são bolsistas de produtividade – CNPq.

## 6 REFERÊNCIAS

BAKURADZE, T. et al. Antioxidant-rich coffee reduces DNA damage, elevates glutathione status and contributes to weight control: results from intervention study. **Molecular Nutrition and Food Research**, Cleveland, v. 55, p. 793-797, 2011.

BORÉM, A. A aplicação dos marcadores moleculares no melhoramento. In: _____. **Marcadores moleculares**. 2. ed. Viçosa, MG: UFV, 2009. p. 95-102.

CAMPA, C. et al. **Candidate gene strategy for the study of the chlorogenic acid biosynthesis**. Montpellier: [s.n.], 2004.

CHAGNE, D. et al. QTL candidate gene mapping for polyphenolic composition in Apple fruit. **BMC Plant Biology**, Bethesda, v. 12, n. 12, p. 1-16, Dec. 2012.

COGAN, N. O. et al. Validation od in silico-predicted genic SNPs in white clover (*Trifolium repens* L.), an outbreeding allopolyploid species. **Molecular Genetics Genomics**, Berlin, v. 277, n. 4, p. 89-113, 2007.

CONSELHO DOS EXPORTADORES DE CAFÉ DO BRASIL. Disponível em: <http://www.cecafe.com.br>. Acesso em: 12 ago. 2011.

DAVIS, A. P. et al. Searching for the relatives of Coffea (Rubiaceae, ixoroideae): the circumscription and phylogeny of Coffeeae based on plastid sequence data and morphology. **American Journal of Botany**, Columbus, v. 94, n. 3, p. 313-329, 2007.

EVANS, D. M.; CARDON, L. R. Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps. **Genetics**, Austin, v. 75, p. 687-692, 2004.

EXPASY TRANSLATE TOOLS. Disponível em: <http://www.expasy.ch/tools/dna.html>. Acesso em: 10 fev. 2012.

FARAH, A. et al. Correlation between cup quality and chemical attributes of Brazilian coffee. **Food Chemistry**, Oxford, v. 98, n. 2, p. 373-380, 2006.

KIM, J. G.; BEPPU, K.; KATAOKA, I. Varietal differences in phenolic content and astringency in skin and flesh of hardy kiwifruit resources in Japan. **Scientia Horticulturae**, Amsterdam, v. 120, n. 4, p. 551-554, 2009.

KOCHKO, A. de. et al. Genetic mapping of caffeoyl-coenzyme A 3-O-methyltransferase gene in coffe trees: impact in chlorogenic acid content. **Theoretical and Applied Genetics**, Berlin, v. 107, n. 4, p. 751-756, 2003.

LAI, Z. et al. Identification and mapping of SNPs from EST sunflower. **Theoretical and Applied Genetics**, Berlin, v. 111, n. 8, p. 1532-1544, 2008.

LASHERMES, P. et al. Molecular characterization and origin of the *Coffea arabica* L. genome. **Molecular Genome and Genetics**, Oxford, v. 261, p. 259-266, 1999.

LEROY, T. et al. Improving the quality of African robustas: QTLs for yield-and quality-related traits in *Coffea canephora*. **Tree Genetics & Genomes**, Heidelberg, v. 7, n. 4, p. 781-798, 2011.

LIJAVETZKY, D. et al. High throughput SNP Discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a re-sequencing approach and SNPlex technology. **BMC Genomics**, London, v. 8, n. 424, p. 1-11, 2007.

LIN, C. et al. Coffee and tomato share common gene repertories as revealed by deep sequencing of seeds and cherry transcripts. **Theoretical and Applied Genetics**, Berlin, v. 112, p. 114-130, Sept. 2005.

MARRACCINI, P. et al. RBCS1 expression in coffee: *Coffea* orthologs, *Coffea arabica* homeologs, and expression variability between genotypes and under drought stress. **BMC Plant Biology**, Bethesda, v. 11, n. 85, p. 1-23, May 2011.

MCCARTHY, J. et al. Chlorogenic acid synthesis in coffe: an analysis of CGA content and real-time RT-PCR expression of HCT, HQT, C3H1, and CCoAOMT1 genes during grain development in C. canephora. **Plant Science**, Shannon, v. 172, p. 861-1060, Feb. 2007.

MENIN, B. et al. Identification and mapping of genes related to caffeoylquinic acid synthesis in *Cynara carduntus* L. **Plant Science**, Shannon, v. 179, p. 338-347, 2010.

MONDEGO, J. et al. Brazilian coffee genome project consortium: an EST-based analysis identifies new genes and reveals distinctive gene expression features of *Coffea arabica* and *Coffea canephora*. **BMC Plant Biology**, Bethesda, v. 11, n. 30, p. 1-22, Feb. 2011.

PONCET, V. et al. SSR mining in coffee tree EST databases: potential use of EST-SSR as markers for the *Coffea* genus. **Molecular Genetics and Genomics**, Berlin, v. 276, n. 5, p. 436-449, Nov. 2006.

RAVEL, C. et al. Single-nucleotide polymorphism frequency in a set of selected lines of bread wheat (*Triticum aestivum* L.). **Genome**, Ottawa, v. 49, p. 1131-1139, 2006.

RUPASINGUE, H. P. V. The role of polyphenols in quality postharvesting handling, and processing of fruits. In: GOPINADHAN, P. et al. (Ed.). **Postharvesting biology and technology of fruits, vegetavles, and flowers**. Iowa: Wiley-Blackwell, 2008. p. 482.

SCHMID, K. J. et al. Large-scale identification and analysis of genome-wide single-nucleotide polymorhisms for mapping in *Arabidopsis thaliana*. **Genome Research**, Cold Spring Harbor, v. 13, p. 1250-1257, 2003.

VIDAL, R. et al. A high-throughput data mining of SNPs in Coffea spp ESTs suggests differential homeologous gene expression in the allotetraploid *Coffea arabica*. **Plant Physiology**, Bethesda, v. 154, p. 1053-1066, 2010.

VIEIRA, L. G. E. et al. Brazilian coffee genome project: an EST-based genomic resource. **Brazilian Journal of Plant Physiology**, Piracicaba, v. 18, p. 95-108, Jan./Mar. 2006.