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

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

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

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.

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

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 nucleotídidicas 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.*

1 INTRODUCTION

The coffee tree belongs to the Rubiaceae family and to the genus *Coffea* of which are known aproximately 103 species. Among those species two have higher economic importance, *Coffea arabica* L. and *Coffea canephora* Pierre ex A. Froehner which represent 70% and 30% of the total coffee market, respectively (CONSELHO DOS EXPORTADORES DE CAFÉ DO BRASIL - CECAFE, 2011).

[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, lvieira@iapar.br
[5] Empresa Brasileira de Pesquisa Agropecuária/EMBRAPA Café - Rodovia Celso Garcia Cid Km 375 - Cx P. 481 - 86001-970 Londrina, PR - filipe.pereira@embrapa.br

The species *C. arabica* is originated from Western Africa (Ethiopia) and is an allotetraploid plant (2n=4x=44), and the remaining species are diploid (2n=2x=22) and alogamous. The most likely origin of *C. arabica* was a hibridation of *C. eugenioides* S. Moore and *C. canephora* (DAVIS et al., 2007; LASHERMES et al., 1999).

Brazil is the biggest coffee producer and exporter worldwide, being responsible for 30% of international market (CECAFE, 2011). The coffee beverage quality is the main value aggregation factor, since it enables better market prices and, thus, higher competitiveness. The chemical composition of coffee is one of the factors that determine beverage quality. Its taste and aroma are results of the combined presence of many volatile and non-volatile chemical constituents, among them sugars, caffeine, trigoneline, lipids and CGAs (FARAH et al., 2006). These last are responsible for astringency and interfere on the taste (KIM; BEPPU; KATAOKA, 2009).

CGAs are secondary metabolism compounds and derive from the phenylalanine biosynthesis by the phenylpropanoids (CAMPA et al., 2004). There are two CGAs synthesis ways, first the phenylalanine amonialiase enzyme (PAL) catalyses phenylalanine in cinnamic acid, the cinnamate 4-hydroxylase (C4H) catalyses the second hydrolization of the cinnamic acid into coumaric acid, this is catalyzed by the enzyme 4 couramato CoA ligase (4CL) into p-coumaril CoA. From this point on there can be two different routes. In the first, the p-coumaric acid can be catalyzed into caffeic acid by 4-couramato 3-hydroxilase (C3'H) that, after that, is catalyzed by the 4CL enzyme into Caffeoil CoA, in which is added the quinic acid by the enzyme hidroxicinamoil quinate transferase (CQT), originating the molecules of CGAs. In the second, p-coumaric acid is added to quinic acid molecule by the CQT enzyme which originates p-coumaroil quinic acid, which is catalyzed by the C3'H enzyme into CGAs (RUPASINGUE, 2008) (Figure 1).

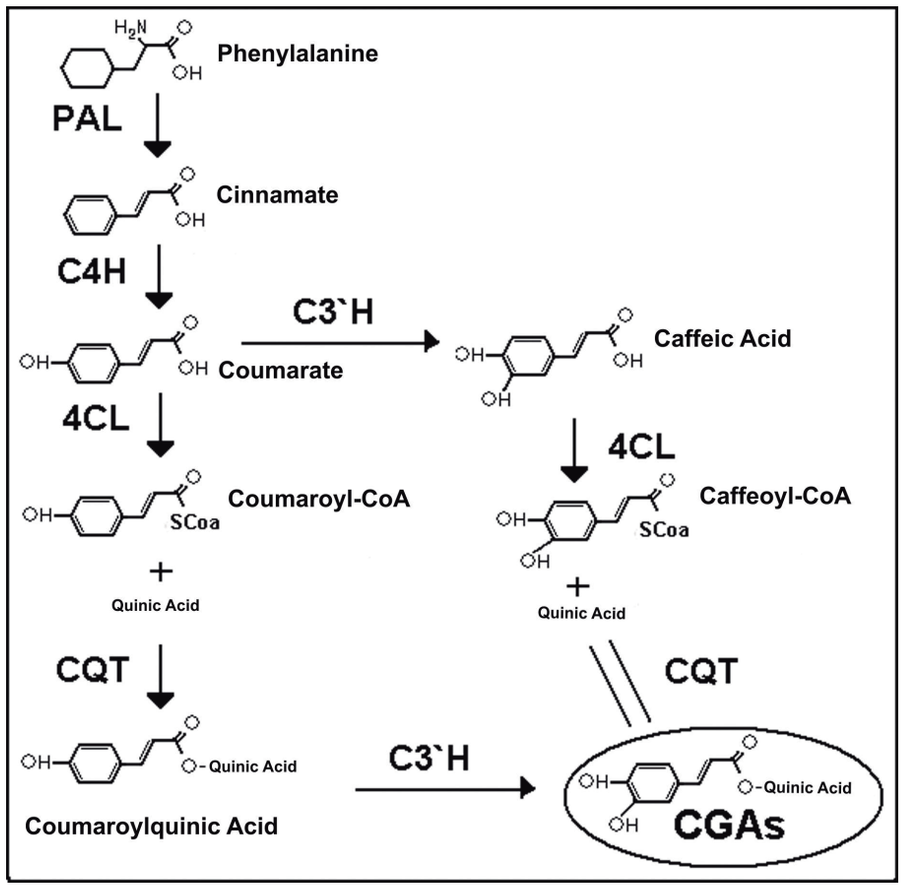

FIGURE 1 - CGAs biosyntheses way and the five enzymes involved.

The study of the enzymes involved in this biosynthesis way is important, for CGAs are the precursors of of lignin synthesis which is involved in protecting the plant's cells against biotic and abiotic stresses (resistance to pathogens and cellular wall permeability in relation to water), besides being fundamental to the plants' rigid structure (KOCHKO et al., 2003; MCCARTHY et al., 2007). The interest in CGAs use in human health is increasing due to its antioxidant properties, antagonist effect to opioids and its capacity to transport glucose in diabetes combat (BAKURADZE et al., 2011).

With the development of molecular biology, new techniques came up to aid genetic enhancement. It is the case of molecular markers which can indicate precisely the individuals' genetic variability (BORÉM, 2009). These have countless advantages in relation to morphologic markers, for they do not depend on environmental conditions and plants' physiologic stage, besides allowing an early genotype identification with characteristics of interest (EVANS; CARDON, 2004). The mapping of genes related to CGAs has been performed in several plant species such as the apple (CHAGNE et al., 2012) and the artichoke (MENIN et al., 2010).

Amongst the various types of available molecular markers, the most interesting for genetic mapping study are microsatellites (SSRs) and unique basis polimorphisms (SNPs). These possess a higher frequency within the organisms' genome and allow the identification and mapping of genes that control features of agricultural interest, interesting factor to reduce time and cost for plants' classic enhancement (BORÉM, 2009).

The use of data from coffee transcriptome, such as EST (express DNA sequences) is allowing studies targeting to better understand the genetic functioning of the plant (MONDEGO et al., 2011; VIDAL et al., 2010). In the *Genbank* (NCBI) besides more than 260.000 sequences from Coffee Genome Project, are also available approximately 47 thousand EST sequences of *canephora* (LIN et al., 2005) and 10 thousand from IRD (Institute de Recherche pour le Developpment) (PONCET et al., 2006).

This work aims the search and selection *in silico* of enzyme sequences involved in CGA biosynthesis, with the objective to identify the number of isoforms for each enzyme, assess the polimorphism level available inside ESTs, analise the origin of sufh polimorphisms (interspecific or intraspecific), calculate their frequencies according to the regions (UTRs or codon) and start an isoform identified mapping within the populations of *Coffea* spp. based in previous analyses using biocomputer tools.

2 MATERIALS AND METHODS

The search for sequences of interest were performed at the Coffee Genome (Genoma Café (http://www.lge.ibi.unicamp.br/cafe) platform through the tools *BlastX* and in *Genbank*/NCBI (http://www.ncbi.nlm.nih.gov/) by *tBlastN* using coding protein sequences for the enzymes involved in CGA biosynthesis previously identified in the genome of *Arabidopsis thaliana* (L.) Heynh. After comparison, were selected only sequences with *e-value* lower than e^{-10}. The sequences found were inserted in the software *Codon Code Aligner* (version 1.6.3) for analysis and *contigs* formation. The same software cleaned the sequences and the quality analyses of the chromatograms through tools called *call base*, *clip ends*, *trim vector* and *find heterozygous* (Phred e Phrap) to exclude low quality regions and discard sequences that correspond to vectors. The alignment parameters of the sequences were the presence of minimum identity percentage of 90% and minimum homology of 20 pairs per base. After the formation of *contigs*, the consensus sequence of those was analised through *BlastX* in *Genbank*/NCBI to validate the protein of interest coding. After confirmation, the reading board of the protein was determined by the analysis performed in online platform EXPASY Translate Tools (2012).

The identification of the isoform number for each enzyme was defined through comparison of the protein sequences of their respective *contigs* using bl2seq (blast two sequence/*Genbank*/NCBI), in which were considered different isoforms when homology and identity percentage showed values lower than 90%.

The detection of polimorphisms was performed based on the quality of chromatograms and only when at least two sequences were detected (minimum of four sequences per contig). Specific oligonucleotides were designed to analise preferably the regions with larger polimorphic frequency observed *in silico*.

3 RESULTS AND DISCUSION

The *in silico* search of polymorphisms for the enzymes PAL, C4H, 4CL, CQT and C3'H and formation of their isoforms selected a total of 426 EST sequences from the Coffee Genome Project database. The total number of ESTs for each of the five CGA enzymes, as well as the amount of contigs formed and the number of isoforms found are described in Table 1.

TABLE 1- Selection oc ESTs and characterization of contigs and singletons found in silico.

ENZYME	ESTs	Contigs*	Singletons	Isoforms (Full Length)	Protein (Partial Length)	
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 formed by, at least, 4 EST sequences

BlastX analyses of the 16 contigs formed in *Genbank* (NCBI) resulted in similarity with some sequences previously noted for coding of proteins related to CGA matebolism (Table 2). Validation with the *Genbank* data is important targeting to confirm the mounting and size of the contigs, as well as their classification within CGAs. All of the isoforms found in this study were deposited in *Genbank* (NCBI). It is interesting to observe that the number of ESTs of the two first enzymes from the phenylpropanoid way, PAL, C4H e 4CL, was three times bigger than the one from the two last enzymes involved in CGA formation. These enzymes respond not only to CGA production, but also to the production of a series of other compounds, such as, for example, flavonoids, salicylic acid and stilbenes.

The total number of EST sequences selected by the *in silico* searches for each contig are represented in Table 3. These sequences were separated according to their respective species of origin by the fact that they influence the polymorphism frequency percentages.

According to the data presented in Table 3, it is possible to notice that, in almost every contig there is a dominance of sequences of *C. arabica*, except for Contig 3 of PAL which showed balance, indicating that isoform must possess a higher expression in *C. canephora*. Such result also is due to the major presence of libraries of *C. arabica* in Coffe Genome Project (Projeto Genoma Café) (VIEIRA et al., 2006).

In the 16 contigs were identified a total of 248 polymorphisms, being 47 inside UTR regions and 201 inside codon regions. Inserted in that last one 63% of polymorphisms (127) correspond to the synonymous tipe (S) and 37% of them (74) to the non synonymous (NS) (Table 4).

A similar percentage was verified for wheat by RAVEL et al. (2006) which found approximately 33% S and 67% NS, demonstrating that most usual mutations do not alter the type of amino acid formed.

The polymorphisms were found in non protein coding regions (5`UTR and 3`UTR), as well as in coding regions. The results highlight the higher SNPs frequency (95%) when compared to indels (5%), as well as the synonym polymorphisms (63%) when compared to the non synonyms (37%) as was observed by SCHMID et al. (2003).

From the detection of polymorphisms the polymorphism presence frequency was calculated in the 3 different sequence regions (5`UTR, protein coding region and 3`UTR) for each 100 base pairs, according to its respective contig. A frequency of 1 polymorphism was found for every 50 base pairs in UTR regions and 1 polymorphism for every 81 base pairs in the protein coding regions.

The research performed *in silico* to know the frequency of polymorphisms (Table 5) according to its location is important, for it is possible to deduce the most likely region to find them in *in vivo* analysis. With such information, it is possible to draw the specific oligonucleotides, prioritizing the regions that show higher polymorphism frequency for their detection. According to LAI et al. (2008), in sunflower EST analyses, the success rate in finding polymorphism *in vivo* reached 72%, when previous *in silico* analyses indicated the best regions for oligonucleotide drawing, against 37% success when they were randomly drawn.

The polymorphism frequency data showed that, in most of the times, it is found in the non protein coding regions (3`UTR e 5`UTR), as reported by MONDEGO et al. (2011).

TABLE 2 - Result of BlastX of the 16 contigs compared against GenBank (NCBI).

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

TABLE 3 - Table containing the total number of sequences and distribution if species of *Coffea*.

Enzyme	S^a Total	S^a (Cc)	S^a (Ca)	S^a (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

S^a = Total sequence number; Cc = *Coffea canephora*; Ca = *Coffea arabica*; Cr= *Coffea racemosa* Ruiz & Pav.

TABLE 4 - Quantity ant types of polymorphisms found.

Enzyme	Number of total polymorphisms	SNPs	Indels	Region 5'UTR	Coding Region		Region 3'UTR
					Synonyms	Non Synonyms	
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

TABLE 5 - Polymorphism frequency according to their locations.

Gene	Frequency 5'UTR/100 pb	Frequency Coding/100 pb	Frequency 3'UTR/100 pb	Average frequency
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= regions with less than four sequences.

The large frequency of SNPs found in all of the studied enzymes confirms the data reported by LIJAVETZKY et al. (2007) that states that SNPs are the most common polymorphisms in grape genome. The total average SNPs found was of 1.4 polymorphism for every 100 base pairs, a relatively large amount when compared to results obtained by VIDAL et al. (2010) who observed 0,39 polymorphisms, however this is possibly due to the fact that 5'UTR region of C3'H_C25 elevated the observed average. Another study that confirms the larger frequency of SNPs when compared to microsatellites (SSRs) was performed by PONCET et al. (2006) who found one SSR for every 7730bp in ESTs of *C. canephora*, number almost 5.5 times lower than the SNPs detected by this work.

According to the results obtained in detection of polymorphisms and proper identification, one can deduce the existence of interspecies allelic differences (*C. arabica* x *C. canephora*, *C. arabica* x *C. racemosa* and *C. canephora* x *C. racemosa*) as well as intraspecies (*C. arabica* x *C. arabica*, *C. canephora x C. canephora* and *C. racemosa* x *C. racemosa*). The types of SNP polymorphisms detected in *in silico* analyses for the five enzymes are in Table 6.

The inter (E) and intra (I) group polymorphism analyses resulted in finding 71% polymorphisms within C. arabica, 18% within C. canephora, 3.9% between C. arabica and C. canephora, 1.4% between C. arabica and C. eugenioides and 5.6% between C. canephora and C. eugenioides. Such data confirm the ones found by VIDAL et al. (2010), whose study found 81% of polymorphisms within C. arabica, 17% within C. canephora, 4% between C. arabica and C. canephora, 2.5% between C. arabica and C. eugenioides.

Also, were classified the polymorphisms between species (31 polymorphisms) and within them (253 polymorphisms). Within species was observed a larger number of polymorphisms in C. arabica (71%), followed by C. canephora (18,3%). Most SNPs found were located within species C. arabica, due to the larger number of sequences within this species, in most contigs.

Within the species C. arabica it was possible to separate the sequences in 2 different genomic groups (subgenomes), one of them is similar to the group of C. canephora and the other probably is similar to C. eugenioides. Such is possibly due to the fact that C. arabica is a hibrid between C. canephora and C. eugenioides (Figure 2).

TABLE 6 - Types of polymorphisms found *in silico* for the CGA enzymes.

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= no sequences of *C. racemosa.* NA*= there was only one sequence of *C. racemosa.*

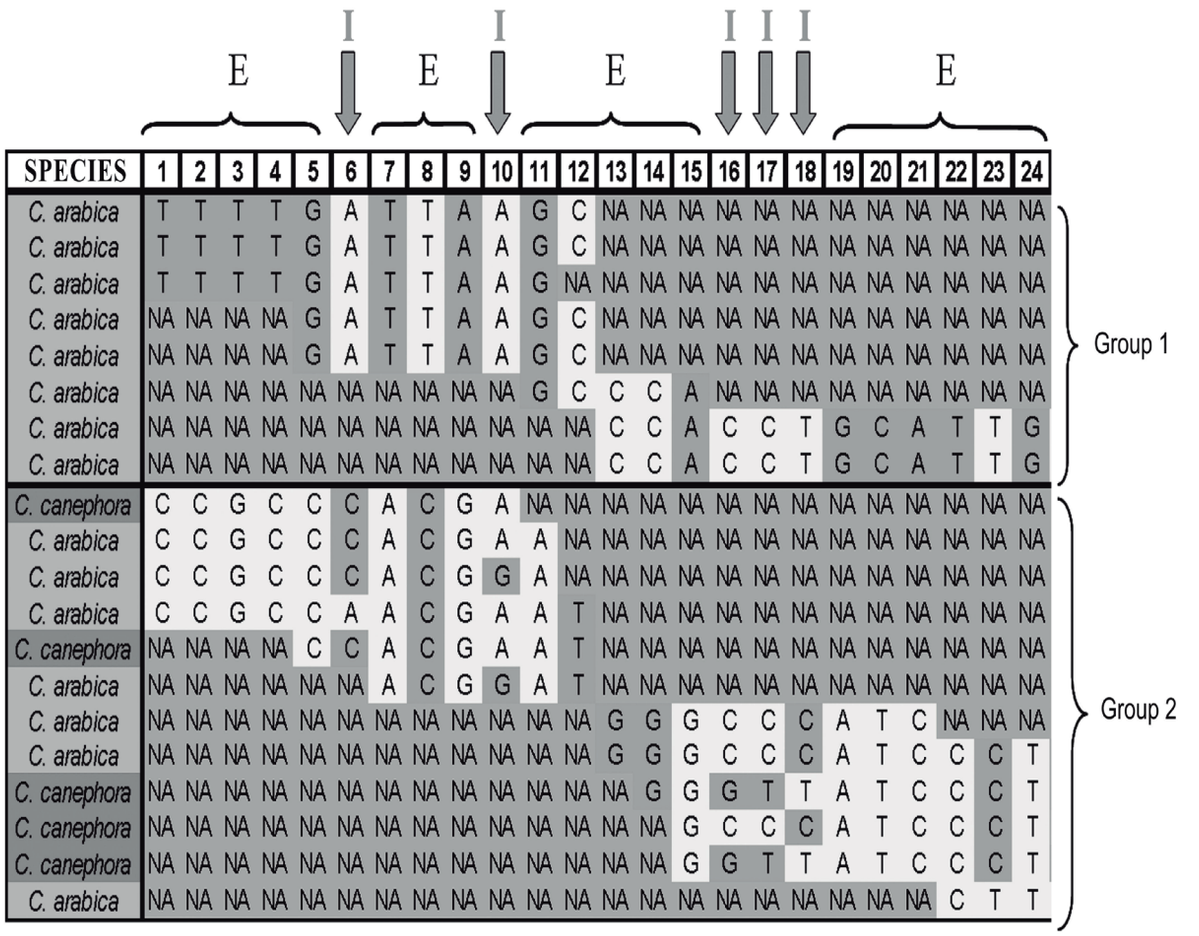

FIGURE 2 - Example of Identification and Characterization of polymorphisms Inta-species and Inter-species (contig4 of 4CL). E: polymorphisms between groups of *Coffea arabica*. I: polymorphisms within groups of *C. arabica*, indicated with arrows. NA: alleles absence. A: adenine. C: cytosine. G: guanine. T: thymine. Group 1: group of *C. arabica* similar to *Coffea eugenioides*. Group 2: group of *C. arabica* similar to *Coffea canephora* and to better observe such homology, some *C. canephora* sequences were inserted in group 2.

The formation of two distinct groups of sequences within *C. arabica*, reinforce the works about the allopolyploid origin of the species from hybridization of two diploid species *C. canephora* and *C. eugenioides* (DAVIS et al., 2007; LASHERMES et al., 1999).

Analysing just the polymorphisms within the subgenomes of *C. arabica*, it was possible to conclude that most of them were located between the two groups (136 polymorphisms) and the smaller part within one of the two formed groups (68 polymorphisms). This SNP separation is important, since it allows characterization studies about the homologous genes expressions in *Coffea arabica*, identifying the differential gene expression from one subgenome compared to another (MARRACCINI et al., 2011; VIDAL et al., 2010). However, aiming to identify the different genotypes and mapping studies, only intraspecific SNPs can be used.

The results generated using ESTs available in public data bases is a basic study that helps and increases the probabilities to develop effective mollecular markers for agricultural interest characteristics, since enables to deduce the nucleotidic diversity expected *in vivo* of individuals of a population, speeding the search process for polymorphic markers. Such factor is essentially important for future mapping and marked assisted selections of economically important perennial species, lowering costs and time for plant's classic enhancement (COGAN et al., 2007).

Based in these *in silico* analyses, some populations of *Coffea* sp. Are being genotyped with primers developed by this work. One of them has already shown positive results and was mapped in *Coffea canephora* (LEROY et al., 2011). For the *Coffea arabica* mapping populations, the analyses are under development.

4 CONCLUSIONS

In silico analyses are effective to identify genes of the mais metabolic ways of coffee tree CGAs. 16 contigs were identified related to five of the main CGA biosynthesis genes. Such genes show a high frequency of polymorphisms of SNP type (95%), with larger distribution in UTR regions (1pol/50pb) when compared to the protein coding regions (1pol/81pb). The *in silico* analysis also identified the presence of two distinct groups of sequences within *C. arabica*, according to their ancestral genomes, allowing to perform homologous gene expression studies. The found genes and polymorphisms have been used in population genotyping studied for *Coffea* sp., to physically and genetically map the CGAs.

5 THANKS

To the Consórcio Pesquisa Café and FINEP/GENOCAFÉ, for the finnancial support. S.T.I. was a Masters Degree intern CNPq; L.F.P.P. and L.G.E.V. are productivity interns – CNPq.

6 REFERENCES

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.