SAMUEL MAZZINGHY ALVARENGA

CARACTERIZAÇÃO FUNCIONAL E IDENTIFICAÇÃO DE SNPS E DE GENES DE RESISTÊNCIA A DOENÇAS DO CAFEEIRO

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Doctor Scientiae*.

VIÇOSA
MINAS GERAIS – BRASIL
2011

SAMUEL MAZZINGHY ALVARENGA

CARACTERIZAÇÃO FUNCIONAL E IDENTIFICAÇÃO DE SNPS E DE GENES DE RESISTÊNCIA A DOENÇAS DO CAFEEIRO

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Doctor Scientiae*.

APROVADA: 18 de julho de 2011.

_____	_____
Pesq. Eveline Teixeira Caixeta (Coorientadora)	Prof. Cláudio Lísias Mafra de Siqueira
_____	_____
Prof. Laércio Zambolim	Pesq. Antônio Carlos Baião de Oliveira

_____

Prof. Ney Sussumu Sakiyama
(Orientador)

“Tudo começa com um sonho. Acredite no seu”

Joslin

À minha querida esposa Ariádine Morgan, dedico.

AGRADECIMENTOS

À Deus, pela vida em abundância, pela saúde, pela força que tem me dado e pela sua presença em TODOS os momentos.

À Universidade Federal de Viçosa e ao Programa de Pós Graduação em Genética e Melhoramento, pela oportunidade de realizar esse curso.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela concessão da bolsa.

Ao Professor Ney Sakiyama, pela orientação, pelas nossas conversas e pelo exemplo de vida.

À Doutora Eveline Caixeta, pela amizade, pelos ensinamentos, por sua paciência, e por ter confiado a mim a execução desse trabalho.

À Doutora Eunize Zambolim, pelos conselhos, pela amizade e por estar sempre pronta a ajudar.

Ao Doutor Antônio Baião, pelas críticas e sugestões no aperfeiçoamento desse trabalho.

Ao Professor Cláudio Mafra, com quem pude aprender muito, pela suas sugestões e colaboração com esse trabalho.

Ao Professor Laércio Zambolim, que contribuiu com seu conhecimento para o enriquecimento das discussões desse trabalho.

Aos amigos do Laboratório de Biotecnologia do Cafeeiro (BIOCAFÉ), onde tive a oportunidade de realizar esse trabalho, pela amizade, brincadeiras e festas, pelo convívio super agradável, pelos ensinamentos e pelo apoio durante esse tempo.

Aos meus grandes amigos Paulo Monteiro, Eduardo Franklin, Felipe Louback, Alisson Xavier e Maíra Freire, pela amizade incondicional, por me fazerem acreditar em mim e pelo apoio nas horas difíceis.

Aos meus amigos da Igreja Presbiteriana de Viçosa (IPV), especialmente Ítalo e Sara Coutinho, Raquel Azeredo, Gino Ceotto, Natan Pimentel, Marcos Bevitori, Zilbinho e Marô pelo convívio, pelas boas influências, pelas orações e por me fazerem sentir especial e importante.

Aos queridos casais Sandra Favero e José de Fátima, Ricardo Monteiro e Lilian Veríssimo, Waldênia Moura e Paulo Lima, José e Silvia Silva, Márcio e Lizzy Veríssimo, Lissânder Dias e Kellen Fonseca, Felipe Stelli e Milena Amaral, Carlos e Shirley Dantas, Paulo Lobato e Fernanda Brandão, Fernando Lee e Letícia Monteiro, pela convivência super agradável e programações divertidas que fizemos juntos.

Aos meus discipuladores Delly Oliveira, Luci Fagundes, Tereza Rocha e Mauro Nacif, pelas conversas, pelos conselhos, pelo carinho e pela amizade.

Ao casal Raphael Klein e Mary Hellen Fabres, pela grande amizade cultivada e ao casal muito querido, Fábio Cintra e Cristiane Vital, pelas palavras consoladoras e pelo renovo da esperança.

Aos meus pais, pelo apoio, pelo compromisso comigo, pelos conselhos, pelas orações e pela atenção. Ao meu irmão Áquila, pelo carinho, preocupação e companheirismo.

À minha esposa, que está sempre ao meu lado e me fazendo acreditar que os sonhos são possíveis, pelo seu grande amor e paciência. Agradeço também aos meus sogros Adenayr Morgan e Josino Marques, e à minha cunhada Harissa, por todo carinho e suporte em palavras renovadoras e orações.

A todos que, de alguma forma, contribuíram para a realização desse trabalho.

Muito obrigado !

BIOGRAFIA

Samuel Mazzinghy Alvarenga, filho de Aurelino Ferreira Alvarenga e Maria de Jesus Mazzinghy Alvarenga, nasceu em Teófilo Otoni, Minas Gerais, no dia cinco de Agosto de 1981.

Em Março de 2000 ingressou no curso de Ciências Biológicas na Universidade Federal de Viçosa, graduando-se em Janeiro de 2005, como Bacharel em Ciências Biológicas.

Em Agosto de 2005 ingressou no Programa de Pós-Graduação, em nível de Mestrado, em Genética e Melhoramento da UFV. Obteve o título de *Magister Scientiae* em Genética e Melhoramento em Julho de 2007.

Em Agosto de 2007 iniciou o Doutorado em Genética e Melhoramento na UFV. Submeteu-se à defesa de tese em Julho de 2011.

SUMÁRIO

RESUMO

ALVARENGA, Samuel Mazzinghy, D.Sc., Universidade Federal de Viçosa, Julho de 2011. **Caracterização funcional e identificação de SNPs e de genes de resistência a doenças do cafeeiro**. Orientador: Ney Sussumu Sakiyama. Coorientadoras: Eveline Teixeira Caixeta e Eunize Maciel Zambolim.

O Projeto Brasileiro do Genoma Café gerou um banco de dados de 200.000 ESTs (*Expressed Sequence Tags*). Estas sequências estão armazenadas no banco de dados do Projeto, o CafEST. Em trabalho anterior, foi realizada análise *in silico* das sequências do CafEST, a qual permitiu a identificação de 14.060 sequências relacionadas com o processo de defesa da planta contra doenças. Dessas 14.060, 1.855 foram sequências cujos produtos preditos eram quitinases. Dada a importância das quitinases, tanto na interação planta-patógeno, como em outros processos da planta, as ESTs de quitinase foram detalhadamente analisadas no presente trabalho. Para isso foi realizada uma caracterização funcional e construído um perfil de expressão *in silico*. A categorização funcional foi feita com o *software* Blast2GO e o perfil de expressão gênica *in silico* foi determinado por meio de análises de *Northern Blot* Virtual. Os resultados da caracterização funcional mostraram a versatilidade das quitinases, que possuem atividade enzimática e estão envolvidas em importantes processos biológicos e funções moleculares nas células da planta. A análise do perfil de expressão *in silico* permitiu verificar que as quitinases estão mais expressas, principalmente, em bibliotecas que apresentam algum componente de estresse. Para selecionar, entre as 14.060 sequências mineradas, aquelas que estão envolvidas com a resistência do cafeeiro à ferrugem, foram desenvolvidos, em trabalho anterior, 40 pares de *primers*. Os *primers* foram testados em 12 cafeeiros resistentes e 12 susceptíveis a *Hemileia vastatrix*, fungo causador da ferrugem. Vinte e nove *primers* resultaram em bandas únicas e bem definidas, sendo que um deles foi polimórfico entre os cafeeiros resistentes e suscetíveis. No presente trabalho, 15 *primers* monomórficos foram escolhidos para análises de polimorfismos de base única (SNPs – *Single Nucleotide Polymorphisms*) nas sequências de DNA entre os 12 cafeeiros resistentes e 12 susceptíveis a *H. vastatrix*. As sequências obtidas foram analisadas com os programas Sequencher® v. 4.10, ORF Finder, BLAST e ClustalW. Dos 15 genes, quatro

não apresentaram nenhum SNP. Cinco genes apresentaram SNPs, mas os polimorfismos foram observados apenas para o clone da espécie *Coffea liberica* var. *dewevrei* (café excelsa). Os outros seis genes foram polimórficos entre os genótipos amostrados. Foram detectados 71 SNPs, sendo que 34 foram transições (47,89%) e 37 transversões (52,11%). A taxa de transições/ transversões foi de 0,9189. Das 71 substituições, 27 ocorreram na primeira posição do códon (38,02%), 15 na segunda posição (21,12%) e 29 na terceira (40,84%). Foram detectadas 11 (15,49%) substituições sinônimas e 60 (84,51%) substituições não sinônimas. A relação de substituições sinônimas/não sinônimas foi de 0,1833. Um alto nível de mutações não sinônimas, como o que foi encontrado neste trabalho, pode ser reflexo de seleção positiva. Um exemplo é o cenário antagonista da coevolução patógeno-hospedeiro, onde os genes do hospedeiro estão engajados numa "corrida armamentista" evolucionária e sob forte pressão de seleção para mudarem e se adaptarem. Em função disso, eles evoluem numa taxa mais rápida que outros genes. Este cenário pode ser aplicado ao presente trabalho, dado o alto nível de modificações não sinônimas detectado e ao fato que os genes analisados atuam, de alguma forma, no processo de defesa do cafeeiro contra patógenos. Foi observada uma frequência de 0,1029 SNPs por amplicon. Nos 119.061 pb analisados detectou-se uma frequência extremamente baixa de 1 SNP a cada 1.676,91pb, ou de 0,0596 SNPs por 100 pb. Pode-se atribuir o baixo nível de polimorfismo observado à baixa diversidade do genoma de *C. arabica*. A análise dos produtos gênicos permitiu verificar que as mutações não sinônimas identificadas não levaram à formação de proteínas preditas distintas. No presente trabalho foi também iniciada a clonagem de um gene potencialmente envolvido com a resistência do cafeeiro a ferrugem. Em trabalho anterior, foi identificado um fragmento de gene de resistência, amplificado pelo *primer* CARF 005, presente nos indivíduos resistentes e ausente nos susceptíveis à ferrugem do cafeeiro. Esse *primer* foi utilizado no presente trabalho para realizar um *screening* de uma biblioteca de BACs (*Bacterial Artificial Chromosome* – Cromossomo Artificial de Bactéria) contendo 56.832 clones Na identificação do(s) clone(s) positivo(s), utilizou-se o método de decomposição de *pools*. Foram identificados dois clones positivos (i. e. dois clones contendo fragmento de gene de resistência amplificado pelo marcador CARF 005). Os dados de sequenciamento dos dois clones identificados poderão fornecer informações importantes sobre a estrutura dos genes de resistência em *Coffea*.

ABSTRACT

ALVARENGA, Samuel Mazzinghy, D.Sc., Universidade Federal de Viçosa, July, 2011. **Functional characterization and identification of SNPs and coffee disease resistance genes**. Adviser: Ney Sussumu Sakiyama. Co-advisers: Eveline Teixeira Caixeta and Eunize Maciel Zambolim.

The Brazilian Coffee Genome Project has generated a database of 200,000 ESTs (Expressed Sequence Tags). These sequences are stored in the Project database, the CafEST. *In silico* analysis of the CafEST sequences was performed in previous work. It allowed the identification of 14,060 sequences related to the process of plant defense against diseases. Of the 14,060, 1,855 were sequences whose predicted products were chitinases. Given the importance of chitinases, both in plant-pathogen interaction, as in other plant processes, the chitinases ESTs were thoroughly analyzed in this study. For that, a functional characterization was carried out and an *in silico* expression profile was built. The functional categorization was accomplished using the Blast2GO software, and the *in silico* gene expression profile was determined by virtual Northern Blot Analysis. The results of functional characterization showed the versatility of chitinases, which have enzymatic activity and are involved in important biological processes and molecular functions in the plant cells. The *in silico* expression profile analysis has shown that chitinases are more expressed especially in libraries that have some component of stress. To select among the 14,060 sequences mined, those which are involved in the resistance of coffee to rust, a set of 40 pairs of primers were designed in previous work. The primers were tested on 24 coffee trees: 12 resistant and 12 susceptible to *Hemileia vastatrix*, fungus that causes leaf rust. Twenty-nine primers resulted in unique and well defined bands, one of which was polymorphic between resistant and susceptible trees. Of these primers, 15 were chosen for Single Nucleotide Polymorphisms (SNPs) analysis in DNA sequences between 12 *H. vastatrix* resistant trees and 12 susceptible ones. The sequences obtained were analyzed with the Sequencher® v.4.10, ORF Finder, BLAST and ClustalW programs. Of the 15 genes, four were monomorphic. Five genes showed polymorphism only for the *Coffea liberica* var. *dewevrei* species clone. The other six genes were polymorphic among the sampled genotypes. Seventy one SNPs were detected, of which 34 were transitions (47.89%) and 37 (52.11%) were transversions. The

rate of transitions/transversions was 0.9189. Of the 71 substitutions, 27 (38.02%) occurred in the first codon position, 15 (21.12%) in the second position and 29 (40.84%) in the third. Eleven (15.49%) synonymous and 60 (84.51%) non-synonymous substitutions were detected. The ratio of synonymous/non-synonymous substitutions was 0.1833. A high level of non-synonymous mutations, as was found in this study, may reflect positive selection. An example is the antagonist scenario of the host-pathogen coevolution, where the host genes are engaged in an "arms race" and under strong evolutionary selection pressure to change and adapt. As a result, they evolve at a faster rate than other genes. This scenario can be applied to this work, given the high level of non-synonymous changes detected and the fact that the analyzed genes somehow act in the defense against coffee pathogens. A frequency of 0.1029 SNPs per amplicon was observed. In the 119,061 bp analyzed, an extremely low frequency of 1 SNP every 1,676.91 bp or 0.0596 SNPs per 100 bp was detected. The low polymorphism level observed can be attributed to the low diversity of the *C. arabica* genome. The gene products analysis has shown that the non-synonymous mutations do not lead to formation of distinct predicted proteins. Also, in the present work, the cloning of a gene potentially involved in resistance to coffee rust was initiated. In previous work, a resistance gene fragment was identified. The fragment, amplified by primer CARF 005, is present in the resistant individuals and absent in the coffee leaf rust susceptible genotypes. This primer was used in this study to perform a screening of a BAC (Bacterial Artificial Chromosome) library containing 56,832 clones. The pools fragmentation method was used to identify the positive clone(s). Two positive clones (i.e. two clones containing the resistance gene fragment amplified by marker CARF 005) were identified. The sequencing data from the two identified clones may provide important information about the structure of disease resistance genes in *Coffea*.

1. INTRODUÇÃO GERAL

1.1. CAFÉ E FERRUGEM

O café é um dos produtos mais tradicionais da agricultura brasileira. As estatísticas de 2010 mostram que o país é líder mundial em produção, exportação e consumo interno (Ministério da Agricultura, Pecuária e Abastecimento - MAPA, 2011). O brasileiro consome, em média, 81 litros de café por ano. Esses dados indicam que, no Brasil, o café é a segunda bebida com maior penetração na população, atrás apenas da água e à frente dos refrigerantes e do leite (Associação Brasileira de Indústria de Café – ABIC, 2011).

Das espécies cultivadas, *Coffea arabica* (café arábica) e *Coffea canephora* (café robusta) são as responsáveis por, respectivamente, 74,15% e 25,85% da produção nacional (CONAB, 2010).

A produção do café é frequentemente reduzida por causa de doenças e pragas. A principal doença na cafeicultura é a ferrugem, causada pelo fungo *Hemileia vastatrix*. A ferrugem ataca as folhas, provocando lesões que levam à morte dos tecidos (Kushalappa e Eskes, 1989). Na ausência de controle químico, a ferrugem pode causar perdas de até 50% na produção (Zambolim *et al*., 2005). A principal forma de controle da ferrugem é por meio de fungicidas (Zambolim *et al*., 2005). Embora potencialmente eficiente, o controle químico é um método oneroso para ser adotado pelos produtores, além de ser fonte de riscos ao meio ambiente. Nesse contexto, o desenvolvimento de cultivares resistentes se torna o melhor método de controle, pois é econômico, eficiente e não causa danos ao meio ambiente.

Várias cultivares de café arábica, portadoras de fatores de resistência à ferrugem, foram disponibilizadas por diferentes instituições brasileiras de pesquisa. Das 105 cultivares de café arábica disponíveis para cultivo comercial,

que constam no Registro Nacional de Cultivares (RNC), do Ministério da Agricultura, Pecuária e Abastecimento (MAPA), 54 são portadoras de fatores de resistência à ferrugem (MAPA, 2010).

1.2. BIOTECNOLOGIA COMO SUPORTE AOS PROGRAMAS DE MELHORAMENTO DO CAFEEIRO

Mesmo quando é sabido que uma característica desejável (fenótipo) é controlada por um gene, o melhoramento clássico consome muito tempo. Quando uma característica é controlada por mais de um gene, o melhoramento é ainda mais difícil e demorado. A introdução de uma nova característica num cultivar de café, usando técnicas de melhoramento convencional, pode levar de 20 a 35 anos antes do lançamento de uma nova cultivar (Ribas *et al*., 2006).

Desta forma, um dos grandes desafios apresentado aos melhoristas atualmente é a diminuição do tempo requerido para se desenvolver novas cultivares. É nesse contexto que a biotecnologia pode oferecer algum suporte. A biotecnologia pode superar fatores que limitam a produção de café, que pelo melhoramento convencional são mais difíceis ou mesmo sem solução. Ela pode acelerar os programas de melhoramento convencional e fornecer aos agricultores, material para plantio livre de doenças. Pode criar plantas resistentes às pragas e doenças, substituindo parcialmente os agrotóxicos que prejudicam o meio-ambiente e a saúde humana (FAO, 2004).

Algumas ferramentas da biotecnologia que podem contribuir para o melhoramento de plantas incluem, entre várias outras, genômica, estudos de expressão gênica, marcadores moleculares, clonagem gênica e transformação.

O recente desenvolvimento da tecnologia genômica tem gerado muitas informações e criado bancos de dados de sequências de DNA, que possibilitam a identificação dos fatores genéticos determinantes e/ou associados com características de interesse agronômico.

A busca de informações biológicas relevantes no contexto genômico pode ser chamada de mineração de dados. Um dos aspectos mais importantes na mineração de dados genômicos é a associação de sequências de DNA com as suas respectivas funções biológicas (anotação). Sabe-se que o genoma de um organismo é constituído de milhares de genes e elementos regulatórios. Esses genes podem interagir com seus produtos, gerando várias combinações

complexas, caracterizando a biologia a nível molecular. A representação da função de uma única proteína nesse contexto é uma tarefa muito complicada, e numa proporção genômica essa tarefa é ainda mais difícil (Thomas *et al*., 2007).

Neste processo de anotação, usuários de bancos de dados biológicos podem perder muito tempo na busca por todas as informações disponíveis sobre uma determinada sequência de DNA/proteína. Este processo torna-se mais difícil a medida que aumentam as variedades de terminologias de uso comum, levando a inibição de uma busca eficiente. Por exemplo, se é realizada uma busca em bancos de dados biológicos por novos alvos para antibióticos, é desejável encontrar nos resultados desta pesquisa todos os produtos gênicos envolvidos na síntese de proteínas bacterianas, e que estes possuam sequências ou estruturas significativamente diferentes das proteínas de humanos. Se um banco de dados descreve tal molécula como sendo “envolvida na tradução”, enquanto outro banco de dados usa a anotação “síntese de proteína”, tornar-se-á muito difícil para o usuário – e ainda mais difícil para um computador – encontrar termos funcionalmente equivalentes.

1.3. ONTOLOGIA GÊNICA E O CONSÓRCIO *GENE ONTOLOGY*

A ontologia gênica é capaz de fornecer uma representação mais simplificada desse sistema biológico tão complexo. A ontologia é um tipo de sistema de organização do conhecimento. Ela consiste de um vocabulário controlado que descreve objetos e as relações entre eles de uma maneira formal (Jermey e Browne, 2004).

A anotação funcional baseada na ontologia gênica permite a categorização de genes em classes funcionais, que podem ser muito úteis para entender o significado fisiológico de grandes quantidades de genes e acessar diferenças funcionais entre subgrupos de sequências. O *Gene Ontology* (GO), desenvolvido pelo Consórcio GO (Ashburner *et al*., 2000), fornece uma plataforma de pesquisa adequada para esse tipo de análise, devido à sua grande cobertura biológica e à sua estrutura em gráficos acíclicos dirigidos (DAG – *Directed Acyclic Graphs*) que permite o entendimento do contexto biológico (Conesa *et al*., 2005). O GO consiste de três domínios de conhecimento: função molecular, processo biológico e componente celular. De uma forma bem geral, esses termos descrevem as funções biológicas do produto de genes, os

processos envolvidos na realização dessas funções, e onde na célula essas funções são tipicamente realizadas (Thomas *et al*., 2007).

O termo “Função Molecular” descreve atividades como “atividade catalítica” ou “atividade de ligação”, que ocorrem em nível molecular. Desta forma, os termos não representam entidades (moléculas ou complexos) que realizam as atividades, e não especifica onde ou quando, ou em qual contexto as ações acontecem. As funções moleculares geralmente correspondem a atividades que podem ser realizadas por produtos gênicos individuais, mas algumas atividades são desempenhadas por complexos de produtos gênicos (Smith *et al*., 2003). Exemplos desses tipos de termos incluem “atividade catalítica”, “atividade transportadora”, “atividade de adenilato ciclase” e “ligação de receptor Toll”.

O termo “Processo Biológico” descreve uma série de eventos realizados por um ou mais conjuntos de funções moleculares (Smith *et al*., 2003). Exemplos de processos biológicos incluem “processo fisiológico celular”, “transdução de sinais”, “processo metabólico de pirimidinas” e “transporte de alfa-glicosídeos”. Pode ser difícil distinguir entre processo biológico e função molecular, mas a regra geral é que um processo tem que ter mais de um passo distinto. Entretanto, um processo biológico não é equivalente a uma via metabólica.

Um “Componente Celular” é apenas um componente da célula, mas com a ressalva de que é parte de uma estrutura maior; pode ser uma estrutura anatômica (ex. “retículo endoplasmático rugoso” e “núcleo”) ou um grupo de produtos gênicos (ex. “ribossomo”, “proteossomo” ou um dímero de proteína).

Os termos do GO podem ser mais gerais ou mais específicos e eles se relacionam por meio de conexões do tipo “is_a” (ou seja, “é uma subclasse de”), ou “part_of” (parte de). Exemplos dessas relações são mostrados na Figura 1.

No caso do processo biológico (Figura 1B), a morte celular programada (“programmed cell death”) é o termo mais específico. Ele se relaciona com um termo menos específico, morte celular (“cell death”) por meio da conexão “is_a”. Ou seja, a morte celular programada é um tipo de morte celular. De modo semelhante, a morte celular é uma subclasse de (“is_a”) morte e de processo celular, e ambos são subclasse de processo biológico, o termo mais geral desta ontologia. Os três principais termos ontológicos, Função Molecular, Processo Biológico e Componente Celular, são os termos “raízes” dos gráficos acíclicos dirigidos, e por isso estão sempre no nível 1. As subclasses subsequentes são numeradas de acordo com a sua especificidade. Quanto mais específico o termo, maior o seu nível ontológico. No exemplo da Figura 1A, a função

molecular “ligação em superfície celular de bactérias gram-positivas” (“Gram-positive bacterial cell surface binding”) está no nível 5 da ontologia. No caso do exemplo do componente celular (Figura 1C), o termo “vesícula de transporte de vírion” (“virion transport vesicle”) pode estar no nível 8, 7, 6, 5 ou 4, dependendo da conexão adotada para estabelecer as relações entre este termo e os menos específicos do gráfico.

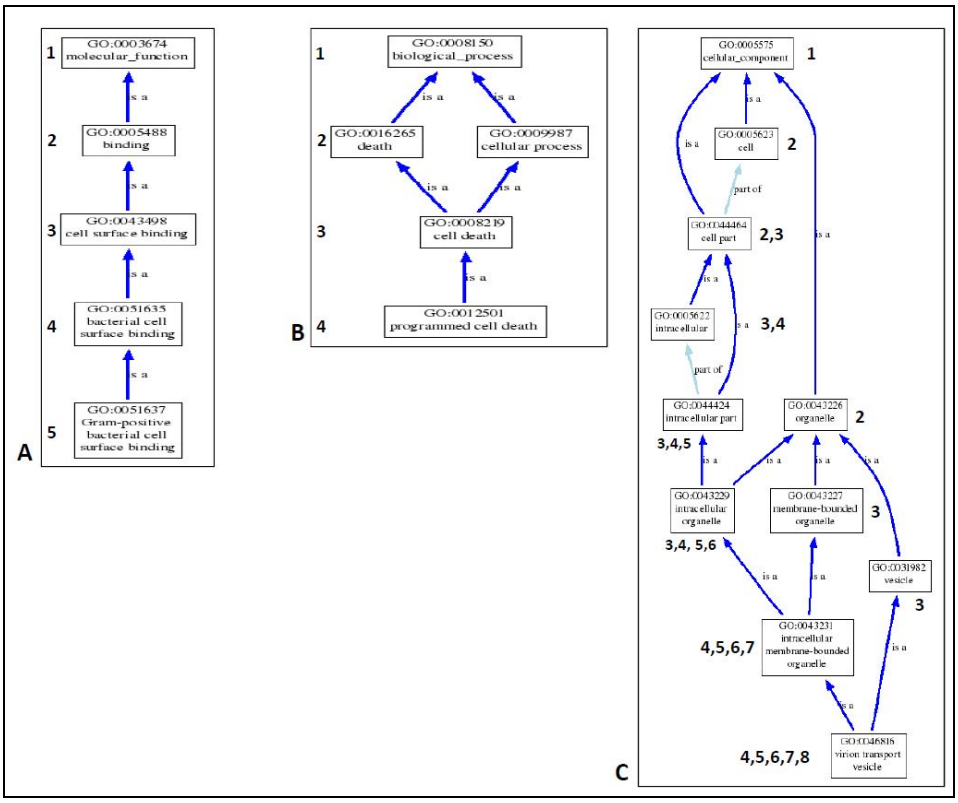


Figura 1 – Relações entre os termos GO. Os termos podem ser mais gerais ou mais específicos. **1A** – Relação entre o termo mais geral “GO:0003674 – Molecular Function”, passando por termos intermediários com conexões “is_a” até o termo mais específico “GO:0051637 – Gram-positive bacterial cell surface binding”. **1B** – Relação entre o termo mais geral “GO:008150 – Biological Process”, passando por termos intermediários com conexões “is_a” até o termo mais específico “GO:0021501 – Programed cell death”. **1C** – Relação entre o termo mais geral “GO:0005575 – Cellular Component”, passando por termos intermediários com conexões “is_a” e “part_of” até o termo mais específico “GO:0046816 – virion transport vesicle”.

O Gene Ontology é utilizado por muitos grupos de pesquisa para uma variedade de tarefas. A Figura 2 mostra quantas vezes o GO foi citado nas mais diferentes disciplinas (http://www.ebi.ac.uk/GOA/users.html em abril de 2011). Na última versão disponível (Janeiro de 2011), o GO continha 32.393 termos indexados, sendo 20.561 de processos biológicos, 9014 de funções moleculares

e 2.818 de componentes celulares (http://www.geneontology.org/GO.downloads .ontology.shtml em abril de 2011).

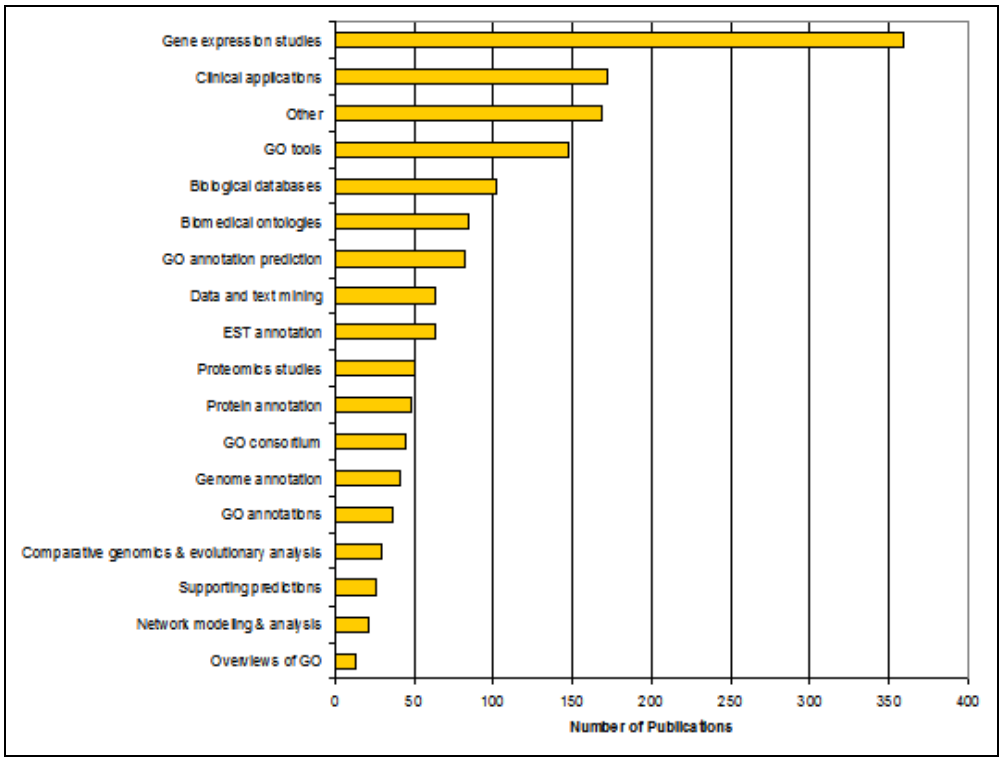


Figura 2 – Frequência de citações de termos GO em diferentes disciplinas.

Todos os termos GO estão associados com uma referência específica que descreve o trabalho ou a análise sob a qual a associação entre o determinado termo GO e o produto gênico está baseada. Cada termo inclui também um “evidence code” (código de evidência) para indicar como a anotação de um termo em particular é suportada. Embora os códigos de evidência reflitam o tipo de trabalho ou análise descrita na referência citada que dá suporte ao termo GO para a associação do produto gênico, eles não são necessariamente uma classificação de tipos de experimentos/analises (www.geneontology.org).

De todos os códigos de evidência disponíveis, apenas “Inferred from Electronic Annotation – IEA” (inferido a partir de anotação) não é atribuído por um curador. Os códigos de evidência atribuídos manualmente estão organizados em quatro categorias gerais: (i) experimental, (ii) análise computacional, (iii) declaração de autores e (iv) declaração de curador. O uso de um código de evidência experimental num termo GO indica que o artigo citado mostra resultados de uma caracterização física de um gene ou produto gênico que deu

suporte à associação deste termo GO. Os códigos de evidência experimentais são:

- Inferred from Experiment (EXP): Inferido a partir de experimento;

- Inferred from Direct Assay (IDA): Inferido a partir de ensaio direto;

- Inferred from Physical Interaction (IPI): Inferido a partir de interação física;

- Inferred from Mutant Phenotype (IMP): Inferido a partir de fenótipo mutante;

- Inferred from Genetic Interaction (IGI): Inferido a partir de interação genética;

- Inferred from Expression Pattern (IEP): Inferido a partir de padrão de expressão (www.geneontology.org).

O uso de códigos de evidência de análises computacionais indica que a anotação é baseada em uma análise *in silico* da sequência gênica e/ou em dados como descritos na referência citada. Os códigos de evidência nessa categoria também indicam um grau variado de curadoria. Os códigos evidência de análises computacionais são:

- Inferred from Sequence or structural Similarity (ISS): Inferido a partir de similaridade de sequência ou de estrutura;

- Inferred from Sequence Orthology (ISO): Inferido a partir de ortologia de sequência;

- Inferred from Sequence (ISA): Inferido a partir de sequência;

- Inferred from Sequence Model (ISM): Inferido a partir de modelo de sequência;

- Inferred from Genomic Context (IGC): Inferido a partir de contexto genômico;

- Inferred from Reviwed Computational Analysis (RCA): Inferido a partir de análises computacionais revisadas (www.geneontology.org).

Os códigos de declaração de autores indicam que a anotação foi feita com base em uma declaração feita pelo(s) autor(es) na referência citada. Os códigos de evidência de declaração de autores são:

- Traceable Author Statement (TAS): Declaração de autor rastreável;

- Non-Traceable Author Statement (NAS): Declaração de autor não rastreável (www.geneontology.org).

O uso de códigos de evidência de declaração de curador indica uma anotação em base no julgamento de um curador que não se encaixa em

nenhuma das outras classificações de códigos de evidência. Os códigos de declaração de curador são:

- Inferred by Curator (IC): Inferido por curador;

- No biological Data Available (ND): Nenhum dado biológico disponível (www.geneontology.org).

A Figura 3 mostra, de forma resumida, em que se baseia cada tipo de código de evidência. Para uma explicação mais detalhada, basta acessar na página do Gene Ontology o *link* para os códigos de evidência (http://www.geneontology.org/GO.evidence.shtml).

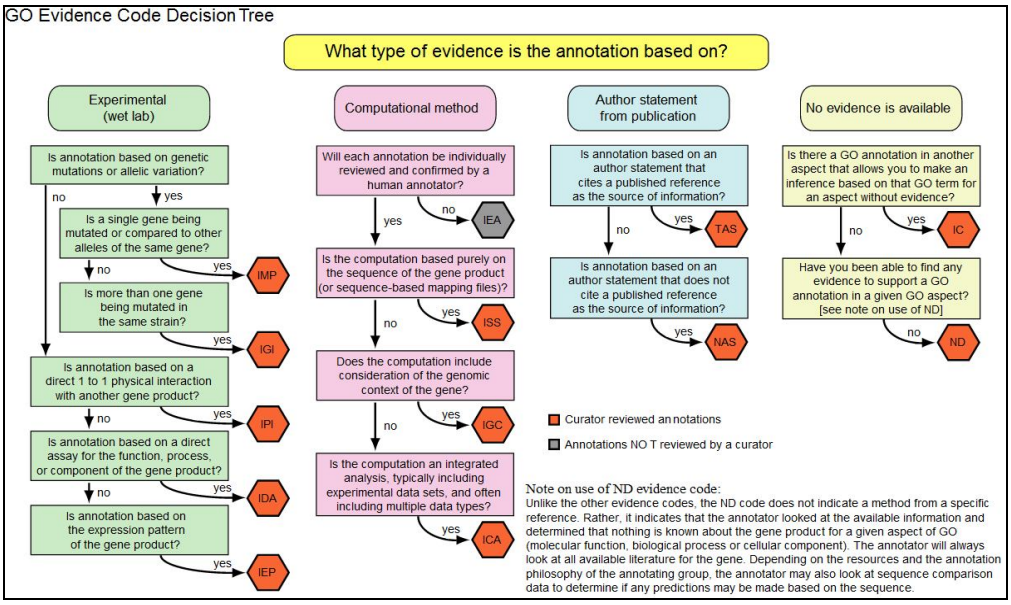


Figura 3 – Fundamentação de cada tipo de código de evidência do GO. Extraída de <http://www.geneontology.org/GO.evidence.tree.shtml>

Os códigos de evidência não relatam a qualidade da anotação. Dentro de cada classe de códigos de evidência, alguns métodos produzem anotações de maior confiabilidade ou maior especificidade que outros métodos. Além disso, a maneira em que uma técnica foi aplicada ou interpretada em um artigo também afeta a qualidade da anotação resultante. Portanto, os códigos de evidência não podem ser usados como uma medida da qualidade da anotação.

1.4. PERFIL DE EXPRESSÃO GÊNICA *IN SILICO*

O estudo da expressão gênica tem demonstrado ser uma importante ferramenta no entendimento dos processos biológicos em nível molecular. Estes

estudos podem, por exemplo, ser utilizados na identificação de redes de genes expressos de importância fundamental no desenvolvimento de determinada estrutura, ou na resposta de um organismo a um estímulo externo. O estudo da expressão gênica diferencial pode contribuir para a caracterização de resistências, pois possibilita a identificação de genes-chave pela comparação da expressão gênica durante o desenvolvimento de uma estrutura sob condições normais e em organismos carregando uma mutação ou submetidos a algum tipo de estresse. O conjunto de dados da expressão diferencial de vários genes sob diferentes condições pode ser usado para reconstituir as vias reguladas por estes genes, predizer onde eles atuam e identificar novos genes associados ao processo (Guimarães *et al*., 2005).

O desenvolvimento de tecnologias de sequenciamento em larga escala resultou em um grande volume de informações sobre sequências completas ou sequências expressas (ESTs) para os mais diversos organismos. O acúmulo destas informações aumentou consideravelmente a demanda por metodologias capazes de analisar simultaneamente um grande numero de sequências, permitindo, desta forma, o estudo de padrões de expressão gênica em diferentes condições biológicas (Andrade *et al*., 2007).

Um tipo de estudo muito importante e que vem sendo adotado por pesquisadores é a análise de expressão gênica *in silico*. Essa análise pode ser usada com o intuito de observar quais genes têm seus níveis de transcrição modificados em determinadas condições biológicas. Pode-se comparar, por exemplo, o nível de expressão de genes em células afetadas com o nível de expressão de genes em células saudáveis de um dado organismo; a medição da modificação da expressão gênica em relação ao tempo, em células durante processos como o ciclo celular, desenvolvimento, respostas a tratamentos externos ou a modificações no ambiente. Com tais informações é possível caracterizar quais genes estão ativos (ou inativos) em um determinado processo biológico de uma forma rápida e ampla. A análise de tais dados com ferramentas computacionais e estatísticas permite identificar genes de interesse no processo biológico em questão. A análise multidimensional de dados de ESTs pode constituir uma nova abordagem para a anotação funcional de genes anônimos e pode contribuir para um entendimento mais global da fisiologia da planta (Ewing *et al*., 1999). O agrupamento de genes por perfil de expressão pode também permitir a identificação de novos elementos regulatórios, uma vez que os genes com perfis correlacionados podem ter elementos regulatórios em comum (DeRisi *et al*., 1997).

1.5. CLUSTERIZAÇÃO HIERÁRQUICA EM DADOS DE EXPRESSÃO GÊNICA *IN SILICO*

Existem duas maneiras confiáveis de se estudar matrizes de expressão gênica: (i) comparar o perfil de expressão de genes entre os genes analisados e (ii) comparar o perfil de expressão dos experimentos (bibliotecas, tratamentos, etc.) entre os ensaios analisados (Southern, 1975). Além disso, se a normalização dos dados permite, é possível uma combinação de ambos. É possível investigar tanto as similaridades como também as diferenças. Se dois genes são similarmente expressos, pode-se inferir que são co-regulados e possivelmente funcionalmente relacionados (D'haeseleer *et al*., 2000). A comparação de experimentos pode fornecer a informação de quais genes são diferencialmente expressos em duas ou mais condições e isso permite estudar, por exemplo, efeitos que vários componentes têm em uma condição investigada.

A clusterização (ou agrupamento) pode ser definida como o processo de separação de elementos em vários grupos com base na sua similaridade (Gilbert *et al*., 2000). Os grupos (ou *clusters*) são formados de modo que as distâncias entre os elementos de um grupo sejam mínimas e as distâncias entre os grupos sejam máximas. Em outras palavras, o objetivo é definir *clusters* que minimizam a variabilidade *intracluster* enquanto maximizam as distâncias *intercluster* (Rodrigues, 2009).

Para medir o quanto um elemento é similar a outro e, assim, identificar se ambos devem estar contidos no mesmo *cluster* ou não, deve ser utilizada uma medida de similaridade (Ochi *et al*., 2004). Em termos de expressão gênica, um valor representando a distância entre dois genes ou experimentos é computado por meio da soma das distâncias entre seus respectivos vetores. Como este valor é normalizado ou como a distância é computada, depende da medida de distância utilizada. Desta forma, quanto menor for a distância entre um par de elementos, maior é a similaridade entre eles (Sturn, 2000).

Existem várias medidas de distância, porém a mais comumente utilizada é a Distância Euclidiana. Ela funciona como medida de dissimilaridade quando utilizada no contexto de técnicas de agrupamento. Ela é definida basicamente como a soma dos quadrados das distâncias de dois elementos (X e Y) num espaço d-dimensional (Figura 4). Ou seja, mede a distância absoluta entre dois pontos. Assim, quanto mais distantes os pontos, menos similares eles são; e quanto mais próximos, mais similares. A Distância Euclidiana leva em

consideração a magnitude dos valores que compõem o ponto, ou seja, atributos com maiores valores (níveis de expressão mais altos) têm maior influência no cálculo da similaridade (Araújo, 2008).

$$Dist_{Eucl}(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}$$

Figura 4 – Fórmula da Distância Euclidiana

Os processos para a solução de problemas de clusterização podem ser classificados, de forma geral, em métodos de particionamento e métodos hierárquicos (Fasulo, 1999).

A clusterização hierárquica (Wen *et al*., 1998; Collins, 1999; Gedina, 2000) cria uma decomposição da base de dados na forma de dendrograma, dividindo-o recursivamente em conjuntos de dados menores. Essa divisão pode ser feita de duas formas: *top-down* e *bottom-up*. Alguns autores chamam essa divisão de divisivos e aglomerativos, respectivamente (Figura 5).

Na abordagem *top-down* (divisivo), o processo inicia com todos os objetos no mesmo grupo, o qual vai sendo dividido sucessivamente até que cada grupo contenha um único elemento. Na forma *bottom-up* (aglomerativo), cada objeto é um grupo e, a cada passo do procedimento, os dois grupos mais próximos (similares) são unidos até que, ao final, exista um único grupo formado por todos os objetos (Doni, 2004).

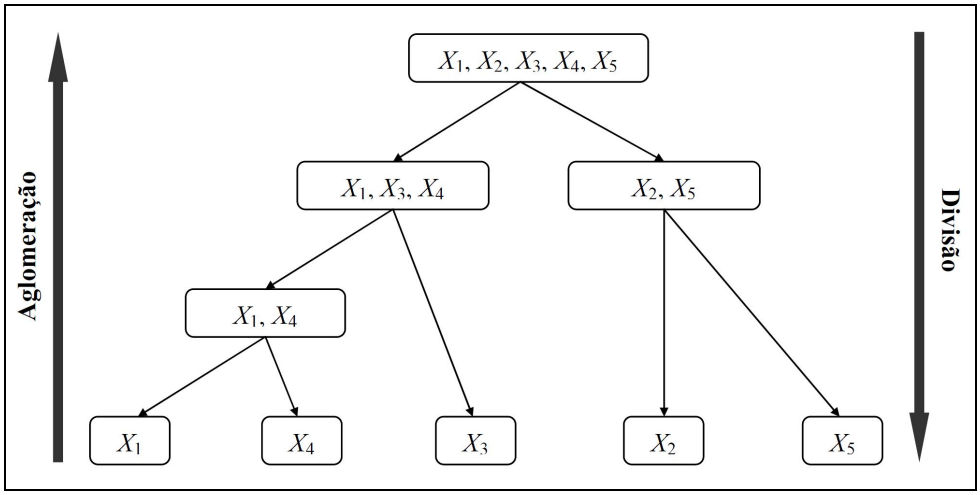


Figura 5 – Exemplo de dendrograma no agrupamento hierárquico.

Existe uma variedade de métodos aglomerativos, que são caracterizados de acordo com o critério utilizado para definir as distâncias entre grupos. Entretanto, a maioria dos métodos parecem ser formulações alternativas de três grandes conceitos de agrupamento aglomerativo (Anderberg, 1973): (i) Métodos de ligação (*single linkage*, *complete linkage*, *average linkage*, *median linkage*); (ii) Métodos de centróide; e (iii) Métodos de soma de erros quadráticos ou variância (método de *Ward*).

Segundo Sturn (2000), a clusterização pelo método de ligação *average linkage* é o que apresenta melhores resultados para dados de expressão gênica. O dendrograma é bem distribuído e os *clusters* são bem visíveis na imagem ordenada. Algumas características desse método incluem (i) menor sensibilidade à ruídos que o os métodos de ligação por vizinho mais próximo e por vizinho mais distante (*single linkage* e *complete linkage*, respectivamente); (ii) apresenta bons resultados tanto para distâncias euclidianas quanto para outras distâncias; e (iii) apresenta uma tendência a formar grupos com número de elementos similares (Kaufman, 1990).

A clusterização hierárquica é a estratégia de clusterização mais usada para análises de expressão gênica na atualidade. A grande vantagem é que poucos parâmetros precisam ser especificados (medida de distância e método de ligação). O resultado é um conjunto reordenado de genes (ou experimentos), onde vetores similares estão próximos entre si no dendrograma e a distância entre os vetores e os *clusters* está codificada no comprimento do ramo do dendrograma. Isto não apenas permite a estimação da similaridade dos genes vizinhos, mas também da distância entre vetores distantes. Isto é muito útil quando se está mais interessado em distâncias do que em similaridades entre duas ou mais condições investigadas (Sturn, 2000).

Outra vantagem do agrupamento hierárquico é a facilidade em lidar com qualquer medida de similaridade utilizada e sua consequente aplicabilidade em qualquer tipo de atributo. As desvantagens relacionam-se à imprecisão do critério de parada e ao fato de que a maioria dos algoritmos desta classe não verifica os grupos formados ao longo de suas execuções. Este último aspecto está relacionado ao fato dos algoritmos para agrupamento hierárquico serem apenas algoritmos construtivos, não permitindo o refinamento de soluções obtidas durante a sua execução (Berkhin, 2002).

1.6. SNPs

As análises *in silico* são necessárias para minerar os dados gerados pelos projetos de sequenciamento. O objetivo dessas análises é de encontrar sequências de genes relacionadas com processos biológicos de interesse. No entanto, o envolvimento do gene no processo biológico de interesse deve ser confirmado por metodologias de genômica funcional. Uma delas é a tecnologia de marcadores moleculares, que permite a detecção de variabilidade do DNA.

Um tipo de marcador molecular que tem sido bastante utilizado a partir do crescimento dos projetos de sequenciamento em larga escala são os SNPs (*Single Nucleotide Polymorphism* - Polimorfismo de Base Única). O SNP é uma variação na sequência do DNA, resultante da diferença de um único nucleotídeo que leva a diferentes alelos entre membros de uma espécie (Zhu e Salmeron, 2007). Diferentemente de outros tipos de polimorfismos, os SNPs são nucleotídeos conhecidos e em posições definidas, que geralmente são detectados por meio de comparação de sequências de DNA.

Os SNPs são classificados de acordo com o tipo de variação de nucleotídeo em transições, purina-purina (A/G) ou pirimidina-pirimidina (C/T), e transversões, purina-pirimidina ou pirimidina-purina (A/C, A/T, G/C, G/T) (Brookes, 1999). Apesar do número de possibilidades de ocorrer variações do tipo transversões ser o dobro de transições, o contrário tem sido observado por vários pesquisadores (Picoult-Newberg *et al*., 1999; Smith *et al*., 2001; Cheng *et al*., 2004). Uma provável explicação, geralmente aceita para DNA de eucariotos, é o processo espontâneo de desaminação da 5-metilcitosina (5mC) para timidina (T) (5’ CpG 3’, onde p é a ligação fosfodiéster 3’ para 5’ entre nucleotídeos adjacentes), dando origem a um maior número de transições de C > T (G > A, na fita reversa) (Cooper e Krawczak 1989; Fryxell e Moon 2004).

Em uma posição específica da molécula de DNA, teoricamente podem existir quatro nucleotídeos possíveis. Mas na realidade, apenas duas dessas quatro possibilidades tem sido observadas em estudos de populações. Embora a natureza bialélica dos SNPs os torne menos informativos por loco examinado quando comparado com marcadores multialélicos como RFLP ou Microssatélites, isso é superado pela sua abundância, que permite o uso de um maior número de locos (Jehan e Lakhanpaul, 2006).

Os SNPs são o tipo mais comum de polimorfismo de DNA (Li *et al*., 2009) e estão amplamente distribuídos nos genomas (Halushka *et al*., 1999), embora

estudos mostrem que a ocorrência e a distribuição varia muito entre espécies. O genoma do milho apresenta 1 SNP por 60-120pb (Ching *et al*., 2002), enquanto em humanos é estimado 1 SNP por 1000pb (Sachidanandam *et al*., 2001). A frequência da distribuição de SNPs dentro do genoma também varia. Na maioria dos organismos estudados até hoje, os SNPs são mais predominantes nas regiões não codificantes do genoma (Soleimani *et al*., 2003).

As mutações presentes em regiões codificadoras são classificadas em sinônimas ou não-sinônimas. Mutações não-sinônimas resultam em uma alteração do aminoácido codificado na sequência protéica, podendo ser conservativas ou não conservativas em função das características dos aminoácidos envolvidos na troca. Nesses casos, a presença do polimorfismo pode levar a uma mudança estrutural da proteína codificada e, consequentemente, a uma possível alteração da sua função (Stitziel *et al*., 2004). Por outro lado, mutações do tipo sinônimas são aquelas nas quais a presença do polimorfismo não causa alteração do aminoácido codificado. Embora as mutações sinônimas não alterem a sequência protéica, elas podem modificar a estrutura e a estabilidade do RNA mensageiro e, consequentemente, afetar a quantidade de proteína produzida (Simões, 2005). Apesar de não haver muitos relatos, sabe-se também que esse tipo de mutação tem o potencial de criar um sítio de *splicing* que pode resultar em modificações fenotípicas (Richard e Beckman, 1995).

A presença de SNPs no genoma pode acarretar consequências no modo como o genoma é expresso, podendo causar *splicing* alternativo do mRNA, alterações no padrão de expressão de genes, geração ou supressão de códons de terminação ou poliadenilação na molécula de RNA mensageiro e alteração nos códons de iniciação de tradução (Guimarães e Costa, 2002).

Após estudos confirmatórios, os SNPs podem ser usados para identificar genes que estão associados à severidade de doenças, resistência a drogas, e na identificação de fenótipos de interesse por genética de associação/ desequilíbrio de ligação (Sachidanandam *et al*., 2001, Bader, 2001). Os SNPs informativos poderão ser usados para seleção assistida (Pavy *et al*., 2006) e incorporados em programas de melhoramento de plantas e animais (Du *et al*., 2009). Além disso, os SNPs também podem ser úteis na construção de mapas genéticos de alta resolução, diagnóstico genético, análises filogenéticas (Rafalski 2002) ou em estudos de história e genética de populações (Halushka *et al*., 1999, Weiss 1998; Evans e Relling 1999; Stephens *et al*., 2001).

Em café, estudos envolvendo SNPs ainda são muito raros. A diversidade de SNPs foi analisada em alguns genes específicos envolvidos no metabolismo de sacarose e de diterpenos, importantes compostos relacionados à qualidade de bebida (Leroy *et al*., 2006). Os resultados indicaram quais genes foram submetidos a seleção natural ou artificial, sugerindo os melhores genes candidatos.

Foram encontrados SNPs também em genes envolvidos nas características químicas do grão de café (Lannes *et al*., 2007). O estudo dessas características é de grande importância para os programas de melhoramento que têm como objetivo a qualidade de bebida. Neste trabalho, a estratégia *in silico* foi complementar à estratégia *in vivo*, permitindo uma avaliação geral dos níveis de polimorfismo dos genes numa larga escala em todo o genoma com um custo relativamente baixo.

Yanagui *et al*. (2010) avaliou a frequência de polimorfismos em genótipos de *C. arabica, C. canephora, C. eugenioides, C racemosa* e *Psilanthus bengalensis*. Neste trabalho também foram analisados oito genes envolvidos na biossíntese de diterpenos e açúcares, bem como um gene mitocondrial e um cloroplastídico, que permitiram também a inferência sobre a história evolutiva dos genes analisados. Aproximadamente 7,6 kb foram explorados para identificação de 465 polimorfismos incluindo 416 SNPs, 18 INDELs e 31 SSR. Uma frequência de 6,1 SNP foi observada a cada 100 pb. Quando todos os polimorfismos foram considerados, verificou-se que 110 correspondem a diferenças entre *Psilanthus* e *Coffea* e 360 correspondem a diferenças entre as espécies de *Coffea*; destes 266 são intra ou inter específicos para *C. canephora* e *C. eugenioides*, espécies ancestrais de *C. arabica*.

Em um estudo recente, a análise de SNPs em larga escala de ESTs de *Coffea* sugeriu uma expressão gênica homeóloga diferencial em *C. arabica* (Vidal *et al*., 2010). As análises permitiram a identificação de diferentes alelos de *C. canephora* e *C. eugenioides* que estão presentes em *C. arabica*. De acordo com essas análises, cerca de 55% das sequências de *C. arabica* são derivadas do genoma de *C. eugenioides* e 45% de *C. canephora*. Além disso, observou-se também que o genoma de *C. eugenioides* contribui principalmente para genes relacionados a metabolismo basal, enquanto que os genes de *C. canephora* estão envolvidos com sinais de tradução de regulação da expressão gênica.

Em outro trabalho, 61 SNPs foram identificados *in silico* a partir de 5.371 ESTs de cafeeiro (Zarate *et al*., 2010). Dezesseis desses SNPs foram validados e dois deles foram polimórficos em genótipos de *C. arabica*. O estudo realizado

por aqueles pesquisadores destacou também as dificuldades em encontrar polimorfismos em espécies de *C. arábica*.

1.7. CLONAGEM GÊNICA E TRANSFORMAÇÃO

A clonagem gênica é um processo no qual um gene de interesse é localizado no DNA e copiado (clonado) múltiplas vezes. Para localizar o gene é necessário construir bibliotecas para catalogar o DNA do organismo. O gene de interesse é então selecionado a partir dessa biblioteca. A etapa seguinte à clonagem é a introdução do gene de interesse em um organismo que não o contém.

Quando um gene é introduzido por retrocruzamento e seleção, não é apenas o gene que é introduzido, mas também as regiões que o flanqueiam (arraste gênico - *linkage drag*). Essas regiões podem carregar genes adicionais, sendo a maioria ainda desconhecidos. Entre esses genes carregados com o arraste gênico pode haver alguns que controlam a síntese de componentes potencialmente nocivos, ou ainda, codificar características agronomicamente indesejáveis. Uma vantagem da biotecnologia sobre o melhoramento clássico é a possibilidade de conhecer o(s) gene(s) a ser(em) inserido(s) (Gepts, 2002).

A transgenia é uma das tecnologias que pode ser utilizada nos programas de melhoramento de plantas. Essa tecnologia oferece duas oportunidades aos melhoristas. Uma é a introdução de uma nova variação genética que não se encontra disponível no germoplasma do programa de melhoramento, e a outra é a criação de fenótipos desejados a partir de genes conhecidos (Zhong, 2001).

O potencial da tecnologia de transformação genética no melhoramento de culturas ainda caminha a passos curtos. Alguns dos poucos casos de sucesso incluem a comercialização de variedades e híbridos com novas características transgênicas como resistência a doenças e a insetos (During, 1996; Jouanin *et al*.,1998) e tolerância a herbicidas (Tsaftaris, 1996).

Em café, vários trabalhos envolvendo clonagem gênica e transformação já foram realizados. Entre eles destacam-se o gene que codifica a subunidade menor da rubisco (Marraccini *et al*., 2003), os três genes envolvidos na síntese de cafeína (Uefuji *et al*., 2003) e a obtenção de café descafeinado (Ogita *et al*., 2003).

2. REFERÊNCIAS BIBLIOGRÁFICAS

ABIC - Associação Brasileira de Indústria de Café. **Indicadores da Indústria de Café no Brasil**. 2010. <http://www.abic.com.br/estatisticas.html>. Acessado em 04 de Julho de 2011.

Anderberg MR (1973) **Cluster analysis for applications**. New York: Academic Press, 1973.

Andrade RV, Mehta A, Guimarães PM, Silva FR, Sá MFG, Saraiva MAP, Fragoso R, Brasileiro ACM (2007) Construção e validação de membranas de macroarranjo de DNA para o estudo da interação planta-nematóide. **Boletim de Pesquisa e Desenvolvimento n. 204**. Embrapa Recursos Genéticos e Biotecnologia. Brasília, DF. 2007.

Araújo DSA (2008) **Algoritmos de agrupamento aplicados a dados de expressão gênica de câncer: um estudo comparativo**. Universidade Federal do Rio Grande do Norte. Dissertação (Mestrado). Sistemas e Computação. 102p.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene Ontology: tool for unification of biology. **Nature**, 25:25-29.

Bader JS (2001) The relative power of SNPs and haplotype as genetic markers for association tests. **Pharmacogenomics**, 2:11-24.

Berkhin P (2002) **Survey of clustering data mining techniques**. Technical Report. Accrue Software.

Brookes AJ (1999) The essence of SNPs. **Gene**, 234:177-186.

Cheng TC, Xia QY, Qian JF, Liu C, Lin Y, Zha XF, Xiang ZH (2004) Mining single nucleotide polymorphisms from EST data of silkworm, *Bombyx mori*, inbred strain Dazao. **Insect Biochemistry and Molecular Biology**, 34:523-530.

Ching ADA, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Rafalski A (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. **BMC Genetics**, 3:19.

Collins FS (1999) Microarrays and macroconsequences. **Nature Genetics**, 21:(1 Suppl): 2.

CONAB – Companhia Nacional de Abastecimento. **Acompanhamento da safra brasileira de café**. Safra 2010, quarta estimativa. 19p. 2010

Conesa A, Götz S, García-Gomez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. **Bioinformatics**, 21:3674-3676.

Cooper DN, Krawczak M (1989) Cytosine methylation and the fate of CpG dinucleotides in vertebrate genome. **Human Genetics**, 83:181-188.

DeRisi JL, Iyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. **Science**, 278:680-686.

D'haeseleer P, Liang S, Somogyi R (2000) Genetic network inference: from co-expression clustering to reverse engineering. **Bioinformatics**, 16:707-26.

Doni MV (2004) **Análise de *Cluster*: Métodos Hierárquicos e de Particionamento.** Monografia, Sistemas de Informação. Faculdade de Computação e Informática, Universidade Presbiteriana Mackenzie. 93p.

Du ZQ, Ciobanu DC, Onteru SK, Gorbach D, Mileham AJ, Jaramillo G, Rothschild MF (2009) A gene-based SNP linkage map for pacific white shrimp, *Litopenaeus vannamei*. **Animal Genetics**, 41:286–294.

During K (1996) Genetic engineering for resistance to bacteria in transgenic plants by introduction of foreign genes. **Molecular Breeding**, 2: 297–305.

Evans WE, Relling MV (1999) Pharmacogenomics: translating functional genomics into rational therapeutics. **Science**, 15:487-491.

Ewing RM, Kahla AB, Poirot O, Lopez F, Audic S, Claverie J (1999) Large-Scale Statistical Analyses of Rice ESTs Reveal Correlated Patterns of Gene Expression. **Genome Research**, 9:950-959.

FAO – Food and Agriculture Organization of the United Nations. 2004. Agriculture Biotechnology: Meeting the needs of the poor? IN: **The state of food and agriculture 2003-2004**. 2004.

Fasulo D (1999) **An Analysis of Recent Work on Clustering Algorithms**. Technical Report, Dept. of Computer Science and Engineering, Univ. of Washington.

Fryxell KJ, Moon WJ (2004) CpG mutation rates in the human genome are highly dependent on local GC content. **Molecular Biology and Evolution**, 22:650-658.

Gedina G (2000), **Praktische Methodenlehre Hintergrund-Material**, lecture notes for “Praktische Methodenlehre”, University of Osnabrueck, Germany.

Gepts P (2002) A comparison between crop domestication, classical plant breeding, and genetic engineering. **Crop Science**, 42: 1780-1790.

Gilbert DR, Schroeder M, van Helden J (2000) Interactive visualization and exploration of relationships between biological objects. **Trends in Biotechnology**, 1:487-494.

Guimarães PE, Costa MC (2002) SNPs: sutis diferenças de um código. **Biotecnologia Ciência & Desenvolvimento**, 26:24-27.

Guimarães PM, Proite K, Leal-Bertioli SC, Bertioli D (2005) Análise *in silico* da expressão gênica diferencial de *Arachis stenosperma* inoculado com *Meloidogyne arenaria*. **Boletim de Pesquisa e Desenvolvimento** n. 85. EMBRAPA Recursos Genéticos e Biotecnologia.

Halushka MK, Fan JB, Bently K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti, A (1999) Patterns of single nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. **Nature Genetics**, 22: 239-247.

Jehan T, Lakhanpaul S (2006) Single Nucleotide Polymorphism (SNP) - Methods and applications in plant genetics: A review. **Indian Journal of Biotechnology**, 5:435-459.

Jermey J, Browne G (2004) **Website Indexing: Enhancing Access to Information within Websites**. Blaxland, NSW: Glenda Brown and Jonathan Jermey.

Jouanin L, Bonade-Bottino M, Girard C, Morrot G, Giband M (1998) Transgenic plants for insect resistance. **Plant Science**, 131: 1–11.

Kaufman L, Rousseeuw PJ (1990) **Finding groups in data: an introduction to cluster analysis**. New York: Wiley.

Kushalappa AC, Eskes AB (1989) **Coffee Rust: Epidemiology, Resistance, and Management**. 1989. Macdonald College/McGill University. CRC Press. ISBN 0849368995, 9780849368998. 360 p. 1989.

Lannes SD, Bouchet S, Ferreira LP, Leroy T, Ivamoto ST, Marracini P, Pereira LFP, Vieira LG, Pot D (2007). Polimorfismos nucleotídicos de genes envolvidos nas características químicas do grão de café. Complementaridade das *estratégias in silico* e *in vivo*. **V Simpósio de Pesquisa dos Cafés do Brasil**, 07-11 de maio, 2007. Águas de Lindóia, SP.

Leroy T, Cubry P, Durand N, Dufour M, De Bellis F, Jourdan I, Vieira LG, Musoli P, Aluka P, Legnate H, Marraccini P, Pot D (2006) *Coffea* spp. and *Coffea canephora* Diversity Evaluated with Microsatellites and SNPs. Lessons from Comparative Analysis. **Proceedings of ASIC Conferences**. 21st Colloquium: Diversity and Breeding. Monpellier, France.

Li F, Kitashiba H, Inaba K, Nishio T (2009) A Brassica rapa Linkage Map of EST-based SNP Markers for Identification of Candidate Genes Controlling Flowering Time and Leaf Morphological Traits. **DNA Research**, 16: 311–323.

MAPA - Ministério da Agricultura, Pecuária e Abastecimento. **Informe Estatístico do Café/junho, 2011**. (http//www.agricultura.gov.br/Vegetal/Cafe/Estatisticas/Café).

Marraccini P, Courjault C, Caillet V, Lausanne F, Lepage B, Rogers WJ, Tessereau S, Deshayes A (2003) Rubisco small subunit of *Coffea arabica*: cDNA sequence, gene cloning and promoter 35 analysis in transgenic tobacco plants. **Plant Physiology and Biochemistry**, 41:17–25. 2003.

Ochi LS, Dias CR, Soares SSF (2004) **Clusterização em Mineração de Dados**. Livro da Escola Regional de Informática Rio de Janeiro – Niterói,RJ. 46p.

Ogita S, Uefuji H, Yamaguchi Y, Koizumi N, Sano H (2003) Producing decaffeinated coffee plants. **Nature**, 423, p. 823. 2003.

Pavy N, Parsons LS, Paule C, MacKay J, Bousquet J (2006) Automated SNP detection from a large collection of white spruce expressed sequences: contributing factors and approaches for the categorization of SNPs. **BMC Genomics**, 7:174.

Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA, Boyce-Jacino M (1999) Mining SNPs from EST databases. **Genome Research**, 9:167-174.

Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. **Current Opinion in Plant Biology**, 94:94-100.

Ribas AF. Pereira LPP, Vieira LGE (2006) Genetic transformation of coffee. 2006. **Brazilian Journal of Plant Physiology**, 18:83-94.

Richard I, Beckman JS (1995) How neutral are synonymous codon mutations? **Nature Genetics**, 10: 259.

Rodrigues FS (2009) **Métodos de agrupamento na análise de dados de expressão gênica**. Dissertação. Estatística. Universidade Federal de São Carlos. São Carlos, SP. 93p.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, *et al*. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. **Nature**, 409: 928-933.

Simões MCM (2005) **Detecção de polimorfismo de base única em etiquetas de sequências expressas de *Schistosoma mansoni***. Dissertação. Biologia Celular e Molecular, Centro de Pesquisas René Rachou/FIOCRUZ, Belo Horizonte, 156 p.

Smith B, Williams J, Schulze-Kremer S (2003) The Ontology of the Gene Ontology. **Proceedings of AMIA Symposium** 2003.

Smith EJ, Shi L, Drummond P, Rodriguez L, Hamilton R, Ramlal R, Smith G, Pierce K, Foster J (2001) Expressed sequence tags for the ckicken genome from a normalized 10-day-old White Leghorn whole embryo cDNA library: 1.DNA sequence characterization and linkage analysis. **Journal of Heredity**, 92:1-8.

Soleimani VD, Baum BR1, Johnson DA (2003) Efficient Validation of Single Nucleotide Polymorphisms in Plants by Allele-Specific PCR, With an Example From Barley. **Plant Molecular Biology Reporter**, 21:281–288.

Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. **Journal of Molecular Biology**, 5:503-517.

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han J, Duan J, Carr JL, Lee MS, Koshy B, Kumar AM, Zhang G, Newell WR, Windemuth A, Xu C, Kalbfleisch TS, Shanner SL, Arnold K, Schulz V, Drysdale CM, Nandabalan K, Judson RS, Ruaño G,Vovis GF (2001) Haplotype variation and linkage disequilibrium in 313 human genes. **Science**, 293:489-493.

Stitziel NO, Binkowski TA, Tseng YY, Kasif S, Liang J (2004) TopoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. **Nucleic Acids Research**, 32:520-522.

Sturn A (2000) **Cluster Analysis for Large Scale Gene Expression Studies**. Dissertação. Institute for Biomedical Engineering, Graz University of Technology, Áustria e The Institute for Genomic Research, Estados Unidos da América. 82 p.

Thomas PD, Mi H, Lewis S (2007) Ontology annotation: mapping genomic regions to biological function. **Current Opinion in Chemical Biology**, 7:4-11.

Tsaftaris A (1996). The development of herbicide-tolerant transgenic crops. **Field Crops Research** 45:115–123.

Uefuji H, Ogita S, Yamaguchi Y, Koizumi N, Sano H (2003) Molecular Cloning and Functional Characterization of Three Distinct *N*methyltransferases Involved in the Caffeine Biosynthetic Pathway in Coffee Plants. 2003. **Plant Physiology**, 132:372–380.

Vidal RO, Mondego JMC, Pot D, Ambrósio AB, Andrade AC, Pereira LFP, Colombo CA, Vieira LGE, Carazzolle MF, Pereira, GAG (2010) 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.

Weiss KM (1998) In search of human variation. **Genome Research**, 8:691-697.

Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, Somogyi R (1998) Large-scale temporal gene expression mapping of central nervous system development. **Proceedings of the National Academy of Science of the United States of America**, 6:334-9.

Yanagui K, Vieira LGE, Pereira LFP, Pot D (2010) Análise da diversidade nucleotídica intra e interespecífica de *Coffea* spp. **Congresso Brasileiro de Genética**, 14 a 17 de setembro de 2010, Guarujá, Brasil . s.l. : s.n., p. 8.

Zambolim L, Zambolim EM, Do Vale FXR, Pereira AA, Sakiyama NS, Caixeta ET (2005) Physiological races of *Hemileia vastatrix* Berk. *et* Br. In Brazil – Physiological variability, current situation and future prospects. In: Zambolim L, Zambolim EM, Várzea VMP (Ed.). **Durable resistance to coffee leaf rust**. Viçosa : UFV, DFP, 2005. p.75-98.

Zarate LA, Cristancho MA, Moncada P (2010) Strategies to develop polymorphic markers for *Coffea arabica* L. **Euphytica**, 173:243–253.

Zhong GY (2001) Genetic issues and pitfalls in transgenic plant breeding. **Euphytica**, 118:137–144.

Zhu T, Salmeron, J (2007) High-definition genome profiling for genetic marker discovery. **Trends in Plant Science**, 12:196-202.

CAPÍTULO 1

CARACTERIZAÇÃO FUNCIONAL E PERFIL DE EXPRESSÃO *IN SILICO* DE QUITINASES DO GENOMA CAFÉ

1. INTRODUÇÃO

O Projeto Brasileiro do Genoma Café (PBGC) gerou um banco de dados de 200.000 ESTs (*Expressed Sequence Tags*) (Vieira *et al*., 2006). Estas sequências estão armazenadas no banco de dados do PBGC, o CafEST (http://www.lge.ibi.unicamp.br/cafe/). Com o intuito de identificar genes determinantes e/ou associados com a resistência do cafeeiro a doenças, Alvarenga *et al*.(2010) realizaram análise *in silico* das sequências do CafEST. Naquele trabalho, foram identificadas 11.300 sequências relacionadas com o processo de defesa da planta contra doenças como, por exemplo, quitinases, proteínas quinase, citocromo P450, proteínas de resistência a doenças, proteínas relacionadas com a patogênese, proteínas com domínio LRR e NBS, proteínas induzidas por hipersensibilidade, dentre outras.

Dessas 11.300 sequências, 1.855 eram quitinases. As quitinases estão presentes em bactérias, fungos, plantas, insetos, crustáceos e alguns vertebrados, podendo apresentar várias funções como metabolismo de quitina, mecanismos de defesa contra patógenos e estresse abiótico, nutrição e parasitismo (Matsumoto, 2006).

As quitinases pertencem às famílias de enzimas glicosil hidrolase 18 e 19, de acordo com a classificação feita por Henrissat e Bairoch (1993). Elas catalisam a hidrólise de quitina, um homopolímero linear de unidades de β-1,4 N-acetil-D-glucosamina (GlcNAc) (Collinge *et al*., 1993) que é comumente encontrado em exoesqueleto de insetos, concha de crustáceos e em parede celular de fungos. A hidrólise enzimática de quitina é realizada por meio de um sistema quitinolítico. Desta forma, as enzimas podem ser endoquitinases, exoquitinases, quitobiases, β-N-acetil hezosaminidases, entre outras classificações.

Dada a importância desta enzima, tanto na interação planta-patógeno, como em outros processos da planta, o objetivo do presente trabalho foi analisar as ESTs de quitinase previamente identificadas, visando a ampliação do conhecimento sobre esta enzima no cafeeiro. Para isso foi realizada a caracterização funcional das sequências e construído o perfil de expressão *in silico*. Uma categoria (ou classe) funcional se refere ao processo biológico, componente celular ou função molecular que o produto de um dado gene pode desempenhar. O perfil de expressão pode mostrar em quais condições a planta tem uma maior atividade dos genes de quitinase. A caracterização funcional,

juntamente com o perfil de expressão, possibilita incorporar informações importantes sobre os genes de quitinase do cafeeiro, contribuindo para o entendimento de como eles atuam na planta.

2. MATERIAL E MÉTODOS

2.1. Sequências de DNA

Foram adotadas nesse trabalho as sequências de quitinases do CafEST identificadas por Alvarenga *et al*. (2010). As 1.855 ESTs encontradas pelo sistema *Gene Projects* (Carazzolle *et al*., 2007) foram clusterizadas, formando 47 EST-*contigs* e 48 *singlets*. Apenas os EST-*contigs* com *e-value* < e-20 e *score* > 100 foram selecionados para a caracterização funcional e a construção do perfil de expressão *in silico*.

2.2. Caracterização funcional

A categorização funcional dos EST-*contigs* selecionados foi feita com o *software* Blast2GO (Conesa *et al*., 2005). O procedimento se iniciou com a realização de um BlastX de todas os EST-*contigs* contra o banco de dados *nr* do NCBI (*National Center for Biotechnology and Information*). O *e-value* máximo do melhor *blast hit* foi ajustado para 1e-10 e o tamanho mínimo do alinhamento (HSP *length*) para 33. Baseando-se no resultado do BlastX, o Blast2GO extrai termos do *Gene Ontology* (GO) para cada EST-*contig*. A distribuição dos termos GO foi analisada ao nível 3 dos Gráficos Acíclicos Dirigidos (DAG – *Directed Acyclic Graphs*). As três categorias de termos designados eletronicamente foram Função Molecular, Processo Biológico e Componente Celular. Posteriormente, os EST-*contigs* foram submetidos ao InterProScan do EBI (*European Bioinformatics Institute*), visando identificar domínios conservados. O InterPro é um banco de dados que acumula informações sobre domínios, motivos e regiões conservadas nas proteínas e famílias de proteínas disponibilizadas por outros bancos de dados que incluem PROSITE, PRINTS, ProDom, Pfam, SMART, TIGRFAMs, PIRSF, UPERFAMILY e PANTHER. A última etapa do Blast2GO consistiu na anotação de ECs (*Enzyme Codes*) e busca de mapas metabólicos do KEGG (*Kyoto Encyclopedia of Genes and Genomes*) para as EST-*contigs*.

2.3. Perfil de expressão *in silico*

O perfil de expressão gênica *in silico* foi determinado por meio de análises de *Northern Blot* Virtual (VNB – *Virtual Northern Blot*) dos EST-*contigs* selecionados. A frequência dos *reads* de cada EST-*contig* para uma determinada biblioteca foi calculada e normalizada. O número de *reads* que compõe cada EST-*contig* foi multiplicado por um fator de normalização de cada biblioteca. O fator de normalização foi obtido pela divisão do número total de *reads* de todas as bibliotecas pelo número de *reads* de cada biblioteca. Foi gerada uma matriz de correlação entre o número de *reads* de cada EST-c*ontig* e as bibliotecas. O padrão de expressão gênica entre EST-*contigs* e bibliotecas foi obtido por meio de clusterização hierárquica, baseado na matriz de distância euclidiana e no método de ligação *average linkage*, usando o *software* Genesis v. 1.7.5 (Sturn *et al*., 2002). Os resultados do *Northen Blot* virtual foram apresentados em forma gráfica de *heat map*.

3. RESULTADOS E DISCUSSÃO

Dos 47 EST-*contigs* de quitinases identificados, 45 tinham *e-value* < e-20 e *score* > 100. Esses foram selecionados para a caracterização funcional e a construção do perfil de expressão *in silico*.

Nos dados analisados foram encontradas apenas enzimas da família glicosil hidrolase 18. As quitinases da família 18 estão presentes em vários organismos, incluindo bactérias, fungos, plantas, insetos, mamíferos e vírus (Matsumoto, 2006). Os membros desta família são sensíveis a alosamidina, um potente inibidor de quitinases. Por outro lado, as quitinases da família 19 parecem ser tolerantes a esse inibidor (Koga *et al*., 1987). Os membros dessa família são encontrados em plantas e em algumas linhagens de *Streptomyces*. Há relatos que essas quitinases são similares à lisozima e à quitosanase, as quais possuem domínio catalítico com alto conteúdo de α-hélice (Matsumoto, 2006).

A maior parte dos EST-*contigs* analisados pertence a classe de endoquitinases (EC:3.2.1.14) ou amilases (EC:3.2.1.0). Esta última tem a propriedade de clivar especificamente ligações O-glicosídicas (Yaldagard *et al*., 2007), enquanto a primeira cliva aleatoriamente as cadeias de quitina, produzindo oligômeros de baixo peso molecular (Matsumoto, 2006). Algumas quitinases de planta também apresentam atividade de lisozima (EC:3.2.1.17)

(Collinge *et al*., 1993), como aconteceu com os EST-*contigs* 38 e 45. Enzimas desta classe hidrolisam ligações glicosídicas entre o ácido N-acetilmurâmico e a N-acetil hexosamina. A análise mostrou ainda que, diferentemente dos demais, o EST-*contig* 29 apresentou atividade de glucanases (EC:3.2.1.39) as quais tem sido relatadas como inibidoras de crescimento fúngico (Van Loon e Van Strien, 1999).

As quitinases são também classificadas de acordo com a sua estrutura primária. São cinco classes no total, porém somente as classes I a IV foram encontradas nos dados analisados.

A busca pelo BLAST mostrou que as EST-*contigs* apresentaram similaridade, principalmente, com sequências de *Coffea arabica* (25) e *Vitis vinifera* (6) (Figura 1). Um estudo sobre a evolução e a composição genômica de *C. canephora* revelou um alto nível de microcolinearidade entre esta espécie e *V. vinifera* (Guyot *et al*., 2009). Com base em análise de sequências de DNA, esses autores relataram ainda um alto nível de conservação entre os genomas de *C. canephora* e outras espécies dicotiledôneas como *Solanum lycopersicon* e *Populus. trichocarpa*.

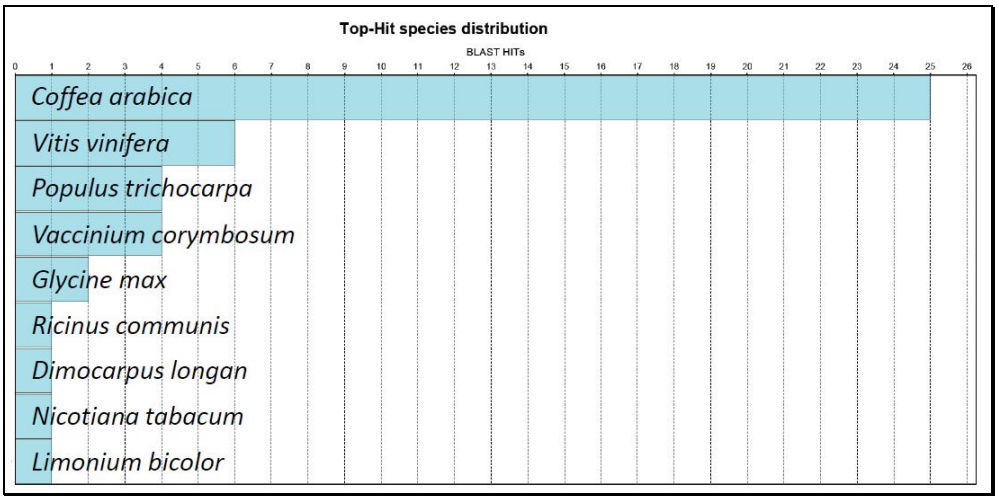


Figura 1 – Distribuição das espécies mais encontradas na busca do BLAST

A função molecular com o maior número de termos associados com as sequências de quitinases foi “atividade de hidrolase” (Figura 2). Este resultado está de acordo com os estudos que demonstram que essas enzimas catalisam a hidrólise de quitina. Por causa dessa propriedade, desde que foram descritas pela primeira vez em 1911, as quitinases são consideradas uma ferramenta de fortalecimento da resposta de defesa de plantas contra patógenos (Sharma *et*

al., 2011). Além disso, o aumento do nível dessas moléculas devido a fatores abióticos e bióticos também contribui para demonstrar o seu papel na resposta de defesa da planta (Gupta *et al*., 2010).

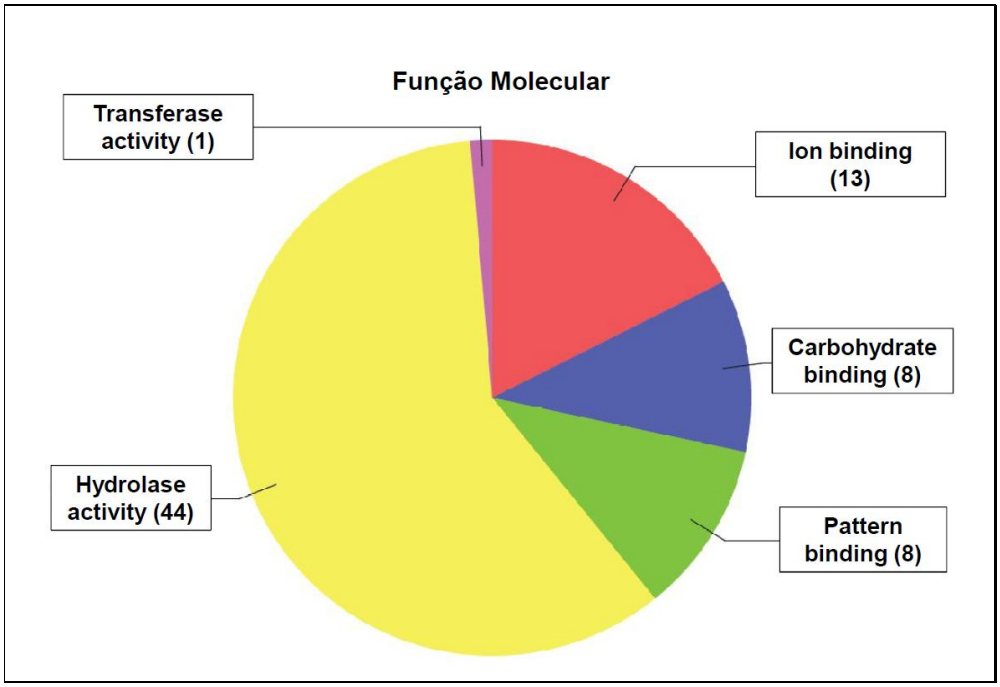


Figura 2 – Distribuição dos termos GO na categoria Função Molecular, nível 3.

A análise da categoria Componente Celular revelou que o termo mais associado com as EST-*contigs* foi parede celular (Figura 3). As quitinases podem ser encontradas em paredes celulares de organismos como os fungos, onde a quitina é um constituinte bastante comum. Neste caso, as quitinases podem ser requisitadas tanto para a morfogênese das paredes celulares como para a lise de quitina, quando há demanda de energia pela célula (Gooday, 1977). Já em plantas, ensaios imunocitoquímicos e estudos de fracionamento celular demonstraram que algumas quitinases da classe I estão localizadas em vacúolos, enquanto outras quitinases estão localizadas apoplasticamente, ou seja, no espaço intercelular (Boller e Métraux, 1988).

O aumento da atividade das quitinases e de outras proteínas relacionadas com a patogênese no processo de defesa da planta envolve uma enorme redistribuição de energia em direção à resposta de defesa contra patógenos (Bolton, 2009). Portanto, a habilidade da célula vegetal em recrutar energia por meio de vias metabólicas que produzem energia (metabolismo primário) é fundamental para a planta. Este cenário foi detectado no resultado da

análise dos processos biológicos. A maior parte dos termos dessa categoria foi "processo metabólico primário" e "processo metabólico de macromoléculas" (Figura 4).

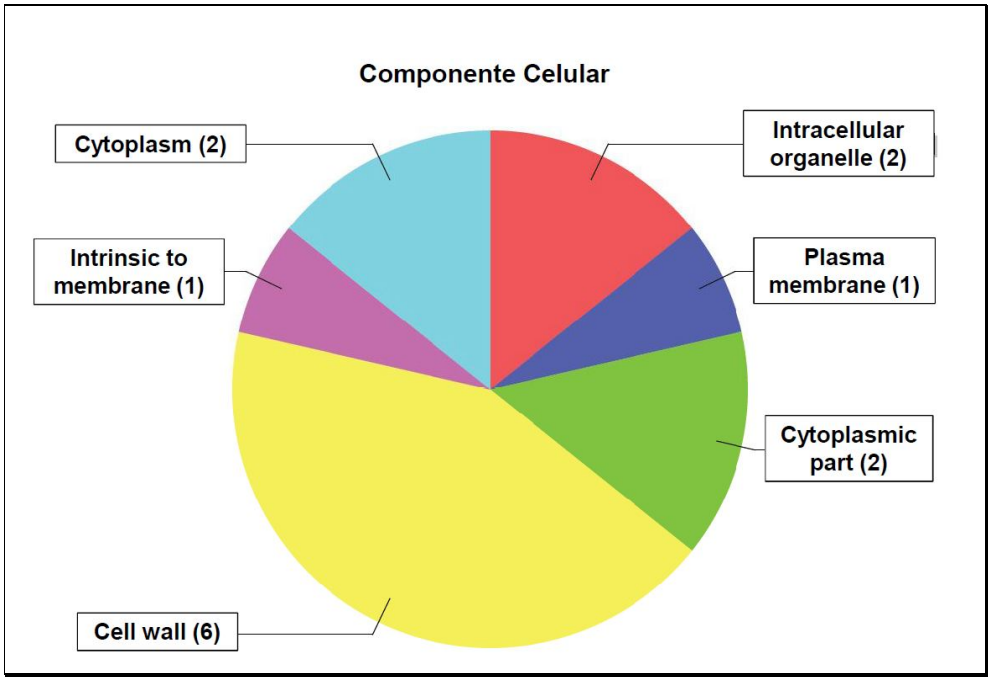


Figura 3 – Distribuição dos termos GO na categoria Componente Celular, nível 3.

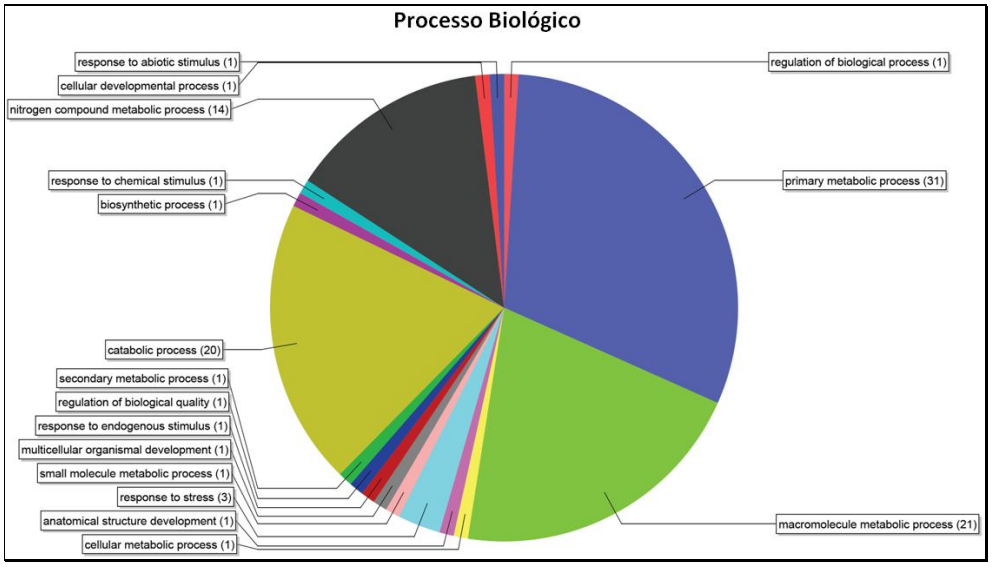


Figura 4 – Distribuição dos termos GO na categoria Processo Biológico, nível 3.

A expressão das quitinases foi mais evidente nas bibliotecas LP1 (plântulas com tratamento de ácido araquidônico), PC1 (linhagem não embriogênica de folhas com indução de 2,4D), RT7 (raízes com BTH), FR2

(botão floral e frutos em estágios diferentes), IA2 (linhagem embriogênica de folhas com indução de 2,4D), FB2 e FB4 (botão floral em diferentes estágios de desenvolvimento), SH1 (folhas de *C. canephora* em estresse hídrico), SH2 (folhas de *C. arabica* em estresse hídrico) e SS1 (condições normais) (Figura 5).

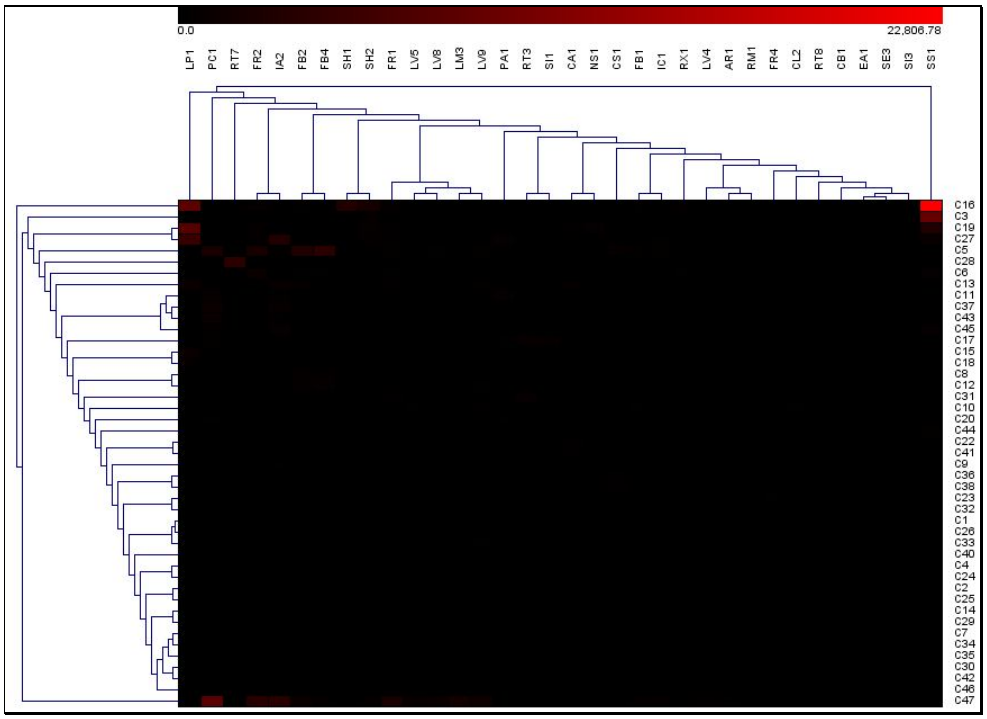


Figura 5 – *Heat map* do perfil de expressão das quitinases nas bibliotecas do CafEST

Todas essas bibliotecas, com exceção da SS1, apresentam algum componente de estresse, o que leva a inferir que as quitinases, de algum modo, podem estar envolvidas neste processo, ou, mais especificamente, no processo de resposta contra esses estresses. O ácido araquidônico (AA), por exemplo (biblioteca LP1), é uma molécula sinalizadora que induz estresse e ativa redes metabólicas de defesa em plantas (Savchenko *et al*., 2010). Alguns estudos demonstraram que o AA é um potente elicitor de morte celular programada e de resposta de defesa em solanáceas (Bostock *et al*., 1981; Garcia-Pineda *et al*., 2004; Knight *et al*., 2001). Foi relatado ainda que o AA induz resistência a vírus em batata e tabaco (Ozeretskovskaya *et al*., 2004; Rozhnova *et al*., 2003) e que induz uma resposta de hipersensibilidade em protoplastos de batata similar àquela induzida por componentes da parede celular do fungo *Phytophtora infestans* (Davis e Currier, 1988). Já o 2,4D (bibliotecas PC1 e IA2) é um herbicida sistêmico seletivo usado no controle de plantas daninhas. Além de

herbicida, o 2,4D também pode atuar na regulação do crescimento da planta (Tomlin, 2006).

A verificação de um alto nível de expressão de quitinases em condições normais (biblioteca SS1) pode ser devido ao fato de que estas enzimas estão entre os genes mais expressos em café (Lin *et al*., 2005). Entretanto, há um outro fator relevante que provavelmente é o que explica melhor este resultado. A biblioteca SS1 é composta por apenas 702 *reads*. Desta forma, quando a frequência dos *reads* é normalizada, o fator de normalização desta biblioteca se torna extremamente alto. Assim, um EST-*contig* que seja formado por poucos, ou apenas um *read* proveniente desta biblioteca, já terá o seu perfil de expressão bastante elevado na biblioteca SS1.

As quitinases são proteínas relacionadas com a patogênese e a sua função na defesa contra patógenos já foi determinada (Yeboah *et al*., 1998). Entretanto, estudos relataram a ação dessas proteínas em outros processos fisiológicos da planta, como regulação do crescimento e do desenvolvimento (Kwon *et al*., 2005; Samac e Shah, 1991; Van der Holst *et al*., 2001), morte celular programada (Passarinho *et al*., 2001), simbiose (Ovstyna *et al*., 2005) e tolerância a vários estresses ambientais (Hamel e Bellemare, 1995; Pinheiro *et al*., 2001; Tateishi *et al*., 2001; Yun *et al*., 1996). As quitinases parecem atuar também no processo de defesa contra estresse hídrico em plantas. Lee *et al*. (2008) investigaram as respostas de proteínas PR na intensidade de estresse hídrico em trevo branco (*Trifolium repens* L.). Foi observado um aumento na atividade de proteínas PR juntamente com uma diminuição no potencial hídrico. A atividade das quitinases se manteve crescente durante todo o período do experimento. Esses resultados indicaram que o aumento da atividade das quitinases e outras proteínas PR nos primeiros dias do experimento pode ser uma expressão de tolerância temporária a estresse hídrico, mas a ativação dessas enzimas durante a fase terminal do estresse pode ser um sintoma induzido pelo estresse hídrico.

Já foi demonstrado, em várias espécies de plantas, que as quitinases são expressas em órgãos como flores, folhas, sementes e raízes (Ancillo *et al*., 1999; Domon *et al*., 2000; Graham e Sticklen, 1994; Passarinho *et al*., 2001; Regalado *et al*., 2000; Samac e Shah, 1991; Takakura *et al*., 2000; Van Hengel *et al*., 1998; Wemmer *et al*., 1994). Sugere-se que algumas dessas quitinases expressas constitutivamente, possivelmente estão envolvidas na defesa da planta, o que dificulta o início da colonização por patógenos (Graham e Sticklen, 1994). As quitinases que são tecido-específicas e não são induzidas por

patógenos ou estresses podem ter papel no crescimento e desenvolvimento da planta (Zhong *et al*., 2002).

Os EST-*contigs* mais expressos nessas bibliotecas foram C16, C3, C19 e C47, todos caracterizados apenas como "*protein*", com função molecular de atividade de hidrolase de componentes o-glicosil e com atividade enzimática de amilase (EC:3.2.1.0). As amilases são uma classe de hidrolases amplamente distribuídas na natureza. Essas enzimas podem clivar especificamente as ligações o-glicosídicas em amido e em outros oligo e polissacarídeos (Yaldagard *et al*., 2007).

4. CONCLUSÕES

A caracterização funcional mostrou a versatilidade dessa proteína, que possui atividade enzimática e está envolvida em importantes processos biológicos e funções moleculares nas células da planta. Os resultados indicaram que as quitinases apresentam atividade de hidrolase, concordando com estudos que demonstram que essa enzima catalisa a hidrólise de quitina. Essa propriedade reforça a importância desta enzima na defesa da planta contra patógenos. O envolvimento dessa enzima com processos de metabolismo primário e de macromoléculas sugere que a célula extrai energia a partir dessas vias metabólicas e a envia em direção à resposta de defesa.

A análise do perfil de expressão *in silico* das quitinases no genoma do cafeeiro permitiu verificar que essas proteínas estão mais expressas, principalmente, em bibliotecas construídas a partir de cafeeiros em situações de estresse. Estes resultados ratificam o envolvimento das quitinases no processo de resposta contra estresses na célula vegetal.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Alvarenga SM, Caixeta ET, Hufnagel B, Thiebaut F, Maciel-Zambolim E, Zambolim L, Sakiyama NS (2010) *In silico* identification of coffee genome expressed sequences potentially associated with resistance to diseases. **Genetics and Molecular Biology**, 33:795-806.

Ancillo G, Witte B, Schmelzer E, Kombrink E (1999) A distinct member of the basic (class I) chitinase gene family in potato is specifically expressed in epidermal cells. **Plant Molecular Biology**, 39:1137–1151.

Boller T, Métraux JP (1988) Extracellular localization of chitinase in cucumber. **Physiological and Molecular Plant Pathology**, 33:11-16.

Bolton MD (2009) Primary Metabolism and Plant Defense - Fuel for the Fire. **Molecular Plant-Microbe Interactions**, 22:487-497.

Bostock RM, Kuc JA, Laine RA (1981) Eicosapentaenoic and arachidonic acids from Phytophthora infestans elicit fungitoxic sesquiterpenes in the potato. **Science**, 212:67–69.

Carazzolle MF, Formighieri EF, Digiampietri LA, Araujo MRR, Costa GGL, Pereira GAG (2007) Gene projects: A genome Web tool for ongoing mining and annotation applied to CitEST. **Genetics and Molecular Biology**, 30:1030-1036.

Collinge DB, Kragh KM, Mikkelsen JD, Nielsen KK, Rasmussen U, Vad K (1993) Plant chitinases. **The Plant Journal**, 3:31-40.

Conesa A, Götz S, García-Gomez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. **Bioinformatics**, 21:3674-3676.

Davis DA, Currier WW (1988) The toxic effect of arachidonic acid and other unsaturated fatty acids on potato tuber cells. **Physiological and Molecular Plant Pathology**, 33:105-114.

Domon J-M, Neutelings G, Roger D, David A, David H (2000) A basic chitinase-like protein secreted by embryogenic tissues of *Pinus caribaea* acts on arabinogalactan proteins extracted from the same cell lines. **Journal of Plant Physiology**, 156:33–39.

Garcia-Pineda E, Castro-Mercado E, Lozoya-Gloria E (2004) Gene expression and enzyme activity of pepper (*Capsicum annuum* L.) ascorbate oxidase during elicitor and wounding stress. **Plant Science**, 166:237–243.

Gooday GW (1977) Biosynthesis of the fungal wall - mechanisms and implications. **Journal of General Microbiology**, 99:1-11.

Graham LS, Sticklen MB (1994) Plant chitinases. **The Canadian Journal of Botany**, 72:1057–1083.

Gupta VK, Misra AK, Gaur RK, Jain PK, Gaur D, Sharma S (2010) Current Status of *Fusarium* Wilt Disease of Guava (*Psidium guajava* L.) in India. **Biotechnology**, 9:176-195.

Guyot R, Mare M, Viader V, Hamon P, Coriton O, Bustamante-Porras J, Poncet V, Campa C, Hamon S, Kochko A (2009) Microcollinearity in an ethylene receptor coding gene region of the *Coffea canephora* genome is extensively conserved with *Vitis vinifera* and other distant dicotyledonoussequenced genomes. **BMC Plant Biology**, 9:22-36.

Hamel F, Bellemare G (1995) Characterization of a class I chitinase gene and of wound-inducible, root and flower-specific chitinase expression in *Brassica napus*. **Biochimica et Biophysica Acta**, 1263:212–220.

Henrissat BI, Bairoch A (1993) New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. **Biochemical Journal**, 293:781-788.

Knight VI, Wang H, Lincoln J-E, Lulai EC, Gilchrist DG, Bostock RM (2001) Hydroperoxides of fatty acids induce programmed cell death in tomato protoplasts. **Physiological and Molecular Plant Pathology**, 59:277–286.

Koga D, Isogai A, Sakuda S, Matsumoto S, Susuki A, Kimura S, Ide A (1987) Specific inhibition of *Bombyx mori* chitinase by allosamidin. **Agricultural and Biological Chemistry**, 51:471-476.

Kwon HK, Yokoyama R, Nishitani K (2005) A proteomic approach to apoplastic proteins involved in cell wall regeneration in protoplasts of Arabidopsis suspension cultured cells. **Plant Cell Physiology**, 46:843–857.

Lee B, Jung W, Lee B, Avice J, Ourry A, Kim T, (2008) Kinetics of drought-induced pathogenesis-related proteins and its physiological significance in white clover leaves. **Physiologia Plantarum**, 132:329-337.

Lin C, Mueller LA, Carthy JM, Crouzillat D, Pétiard V, Tanksley SD (2005) Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts. **Theoretical and Applied Genetics**, 112:114-130.

Matsumoto KS (2006) Fungal chitinases. In: Guevara-González RG, Torres-Pacheco I (eds.) **Advances in Agricultural and Food Biotechnology**, ISBN: 81-7736-269-0. p.289-304.

Ovtsyna AO, Dolgikh EA, Kilanova AS, Tsyganov VE, Borisov AY, Tikhonovich IA, Staehelin C (2005) Nod Factors Induce Nod Factor Cleaving Enzymes in Pea Roots. Genetic and Pharmacological Approaches Indicate Different Activation Mechanisms. **Plant Physiology**, 139:1051–1064.

Ozeretskovskaya OL, Varlamov VP, Vasyukova NI, Chalenko GI, Gerasimova NG, Panina YS (2004) Influence of systemic signal molecules on the rate of spread of the immunizing effect of elicitors over potato tissues. **Applied Biochemistry and Microbiology**, 40:213–216.

Passarinho PA, Van Hengel AJ, Fransz PF, de Vries SC (2001) Expression pattern of the Arabidopsis thaliana AtEP3/AtchitIV endochitinase gene. **Planta**, 212: 556–567.

Pinheiro C, Chaves MM, Ricardo CP (2001) Alterations in carbon and nitrogen metabolism induced by water deficit in the stems and leaves of *Lupinus albus* L. **Journal of Experimental Botany,** 52:1063-1070.

Regalado AP, Pinheiro C, Vidal S, Chaves I, Ricardo CPP, Rodrigues-Pousada C (2000) The *Lupinus albus* class-III chitinase gene, *IF-3*, is constitutively expressed in vegetative organs and developing seeds. **Planta**, 210:543–550.

Rozhnova NA, Gerashchenkov GA, Babosha AV (2003) The effect of arachidonic acid and viral infection on the phytohemagglutinin activity during the development of tobacco acquired resistance. **Russian Journal of Plant Physiology**, 50:661–665.

Samac DA, Shah DM (1991) Developmental and Pathogen-Induced Activation of the Arabidopsis Acidic Chitinase Promoter. **The Plant Cell**, 3:1063–1072.

Savchenko T, Walley JW, Chehab W, Xiao Y, Kaspi R, Pye MF, Mohamed ME, Lazarus CM, Bostock RM, Dehesha K (2010) Arachidonic Acid: An Evolutionarily Conserved Signaling Molecule Modulates Plant Stress Signaling Networks. **The Plant Cell**, 22:3193-3205C.

Sharma N, Sharma KP, Gaur RK, Gupta VK (2011) Role of chitinase in plant defense. **Asian Journal of Biochemistry**, 6:29-37.

Sturn A, Quackenbush J, Trajanoski Z (2002) Genesis: Cluster analysis of microarray data. **Bioinformatics**, 18:207-208.

Takakura Y, Ito T, Saito H, Inoue T, Komari T, Kuwata S (2000) Flower-predominant expression of a gene encoding a novel class I chitinase in rice (*Oryza sativa* L.). **Plant Molecular Biology**, 42: 883–897.

Tateishi Y, Umemura Y, Esaka M (2001) A Basic Class I Chitinase Expression in Winged Bean is Up-regulated by Osmotic Stress. **Bioscience, Biotechnology and Biochemistry**, 65:1663–1668.

Tomlin CDS (2006) **The Pesticide Manual: A World Compendium**, 14th ed.; British Crop Protection Council: Surrey, UK.

Van der Holst PP, Schlaman HR, Spaink HP (2001) Proteins involved in the production and perception of oligosaccharides in relation to plant and animal development. **Current Opinion in Structural Biology**, 11:608-616.

Van Hengel AJ, Guzzo F, Van Kammen A, De Vries SC (1998) Expression pattern of the carrot EP3 endochitinase genes in suspension culture and in developing seeds. **Plant Physiology**, 117: 43–53.

Van Loon LC, Van Strien EA (1999) The families of pathogenesis-related proteins, their activities, and comparative analysis of PR-1 type proteins. **Physiological and Molecular Plant Pathology**, 55:85-97.

Vieira LGE, Andrade AC, Colombo CA, Moraes AHA, Metha A, Oliveira AC, Labate CA, Marino CL, Monteiro-Vitorello CB, Monte DC, *et al.* (2006) Brazilian coffee genome project: An EST-based genomic resource. **Brazilian Journal of Plant Physiology**, 18:95-108.

Wemmer T, Kaufmann H, Kirch H-H, Schneider K, Lottspeich F, Thompson RD (1994) The most abundant soluble basic protein of the stylar transmitting tract in potato (*Solanum tuberosum*L.) is an endochitinase. **Planta**, 194:264–273.

Yaldagard M, Mortazavi SA, Tabatabaie F (2007) The Effectiveness of Ultrasound Treatment on the Germination Stimulation of Barley Seed and its Alpha-Amylase Activity. **World Academy of Science, Engineering and Technology**, 34:154-157.

Yeboah NA, Arahira M, Nong VH, Zhang D, Kadokura K, Watanabe A, Fukazawa C (1998) A class III acidic endochitinase is specifically expressed in the developing seeds of soybean (*Glycine max* [L.] Merr.) **Plant Molecular Biology**, 36**:**407-415.

Yun DJ, D'Urzo MP, Abad L, Takeda S, Salzman R, Chen Z, Lee H, Hasegawa PM, Bressan RA (1996) Novel Osmotically Induced Antifungal Chitinases and Bacterial Expression of an Active Recombinant Isoform. **Plant Physiology**, 111:1219-1225.

Zhong R, Kays SJ, Schroeder BP, Ye Z (2002) Mutation of a Chitinase-Like Gene Causes Ectopic Deposition of Lignin, Aberrant Cell Shapes, and Overproduction of Ethylene. **The Plant Cell**, 14:165-179.

CAPÍTULO 2

IDENTIFICAÇÃO DE POLIMORFISMOS DE BASE ÚNICA EM GENES DE CAFEEIRO ENVOLVIDOS NA DEFESA CONTRA DOENÇAS

1. INTRODUÇÃO

O café é uma das principais commodities agrícolas no mundo, tendo grande importância econômica tanto nos países produtores quanto nos países consumidores. Embora *Coffea arabica* seja a espécie mais cultivada (aproximadamente 70%), esta apresenta uma alta susceptibilidade a patógenos e pragas.

A ferrugem alaranjada do cafeeiro, popularmente conhecida apenas como ferrugem, é causada pelo fungo *Hemileia vastatrix* e é a principal doença que limita a produção de café em quase todos os países produtores no mundo. No Brasil, principal produtor de café arábica, as perdas nas plantações devido a esta doença são estimadas em torno de 30% se medidas de controle não são tomadas.

O uso de cultivares resistentes apresenta-se como a melhor maneira de controlar esta doença por ser viável economicamente e ser ecologicamente correta. *C. canephora* é uma das principais fontes de genes de resistência para os programas de melhoramento do cafeeiro. Entretanto, esta espécie apresenta uma bebida de qualidade inferior. Desta forma, o Híbrido de Timor, proveniente do cruzamento natural entre *C. arabica* e *C. canephora*, constitui-se uma boa opção de fonte de genes, visto reunir as características de qualidade da bebida e resistência a doenças. Além disso, o Híbrido de Timor possui uma alta similaridade genética com *C. arabica*. (Setotaw, 2009). Outro fator importante é a estrutura genética dessas espécies. O Híbrido de Timor é tetraplóide, assim como *C. arabica*, o que viabiliza o cruzamento entre estas espécies, permitindo que arábica receba genes de acessos de Híbrido de Timor.

Estudos sobre a herança da resistência à ferrugem realizados no Centro de Investigações das Ferrugens do Cafeeiro (CIFC), em Portugal, demonstraram que a hipótese de Flor (1971) da teoria gene-a-gene é aplicável à interação cafeeiro-ferrugem onde a resistência completa é condicionada por pelo menos nove genes dominantes de efeito maior (S_H1-S_H9) (Noronha-Wagner e Bettencourt, 1967; Bettencourt e Noronha-Wagner, 1971; Bettencourt *et al*., 1980). Os genes S_H1, S_H2, S_H4 e S_H5 foram encontrados em arábicas puros de origem etíope. O gene S_H3 é considerado como sendo derivado de *Coffea liberica*, enquanto S_H6, S_H7, S_H8 e S_H9 são oriundos de *C. canephora*.

Os fatores S_H oferecem resistência a raças de *H. vastatrix*. Entretanto, quando alguns genes S_H estão suplantados, os cafeeiros podem apresentar resistência incompleta ou parcial a infecções (Eskes, 1989).

Vários métodos têm sido usados para identificar genes de efeito maior que governam características importantes. Entre eles a abordagem de genes candidatos (Candidate Genes - CG) representa uma das estratégias mais adequadas para a rápida identificação e clonagem de genes humanos, animais e de plantas (Pflieger *et al*., 2001). Esta estratégia se baseia na hipótese de que o polimorfismo molecular em genes com função conhecida (isto é, genes candidatos) está relacionado com a variação fenotípica (de Vienne *et al*., 1999; Pflieger *et al*., 2001). Em plantas, vários genes envolvidos com o reconhecimento de patógenos, transdução de sinais e defesa contra patógenos foram isolados de diversas espécies (Lamb *et al*., 1989; Hammond-Kosack e Jones, 1996; Bent, 1996). Assim como os genes candidatos de resistência (Resistance Gene Candidates – RGC), os genes candidatos de defesa (Defense Gene Candidates – DGC) podem ser identificados facilmente por meio de similaridade de sequência em certos domínios funcionais como o sítio de ligação a nucleotídios (Nucleotide Binding Site – NBS) ou os resíduos de leucina repetidos (Leucine-Rich Repeats – LRR) (Meyers *et al*., 1999; Bent, 1996; Pflieger *et al*., 2001).

Partindo dos dados gerados pelo Projeto Brasileiro do Genoma Café, Alvarenga (2007) identificou, por meio de análise *in silico*, cerca de 14.000 sequências de genes RGC e DGC. Visando verificar o envolvimento destes genes com a resistência do cafeeiro à ferrugem foram desenhados 40 *primers* para amplificar algumas das sequências mineradas. Os 40 *primers* foram testados em 12 cafeeiros resistentes e 12 susceptíveis a *H. vastatrix*. Vinte e nove destes *primers* resultaram em bandas únicas e bem definidas, sendo um polimórfico. Em relação aos 28 pares de *primers* que não apresentaram polimorfismo, é possível que os fragmentos possuam polimorfismos de base única (SNPs – *Single Nuclotide Polymorphism*) ao longo da sequência e, por isso, não puderam ser observadas por eletroforese em gel de agarose e/ou poliacrilamida. Os SNPs constituem o tipo de variação mais abundante em genomas eucariotos, e estão presentes tanto em regiões codificadoras como não-codificadoras (Tsui *et al*., 2003).

Os SNPs podem estar associados com características fenotípicas, o que os tornam uma ferramenta muito útil para o desenvolvimento de marcadores moleculares. O catálogo de estudos de associação listava, até julho de 2011,

959 publicações e 4786 SNPs de associações com doenças humanas (Hindorff *et al*.). Desta forma, o objetivo deste trabalho foi verificar a presença de SNPs entre as sequências de 15 genes em 24 genótipos de cafeeiro e avaliar o potencial desse marcador para a espécie.

2. MATERIAL E MÉTODOS

2.1. Material Vegetal

Para a identificação de SNPs nas regiões gênicas pré-selecionadas foram utilizados 24 genótipos de cafeeiros do Banco de Germoplasma da Universidade Federal de Viçosa (UFV)/Empresa de Pesquisa Agropecuária de Minas Gerais (EPAMIG). Desses, 12 foram selecionados por serem utilizados como fontes de resistência a *H. vastatrix*, nos programas de melhoramento (genótipo 13 a 24 da Tabela 1) e 12 genótipos de cafeeiros susceptíveis a esse fungo (genótipos 1 a 12 da Tabela 1).

Tabela 1 – Genótipos utilizados para a identificação de SNPs

Código	Genótipo	Descrição	Grupo fisiológico*	Fator S_H*
1	UFV 570	Bourbon Amarelo da China	E	5
2	UFV 2945	Típica da China	E	5
3	UFV 557-06	Triplóide (*C. arabica* x *C. racemosa*)	?	?
4	UFV 2145-79	Catuaí Vermelho IAC 81	E	5
5	UFV 2154-345	Catuaí Amarelo IAC 86	E	5
6	UFV 2143-193	Catuaí Amarelo IAC 30	E	5
7	UFV 2143-235	Catuaí Amarelo IAC 30	E	5
8	UFV 2148-57	Catuaí Amarelo IAC 64	E	5
9	UFV 2164-193	Mundo Novo IAC 515-3	E	5
10	UFV 2190-100	Mundo Novo IAC 464-18	E	5
11	Caturra 812	Caturra Amarelo	E	5
12	*Coffea liberica* var. *dewevrei*	CIFC 168/12 ou UFV 595	?	?
13	UFV 440-22	Híbrido de Timor	?	?
14	UFV 443-3	Híbrido de Timor	?	?
15	UFV 445-46	Híbrido de Timor	?	?
16	CIFC 832-1	Híbrido de Timor	A	5,6,7,8,9,?
17	CIFC 832-2	Híbrido de Timor	A	5,6,7,8,9,?
18	UFV 438-52	Híbrido de Timor	?	?
19	UFV 446-08	Híbrido de Timor	?	?
20	CIFC 33-1	S 288-23	G	3,5
21	CIFC 1343-269	Híbrido de Timor	R	6
22	CIFC 4106	Híbrido de Timor	A	?
23	CIFC H 420-10	Mundo Novo (CIFC 1535/33) x HW 26/14 (Caturra Vermelho CIFC 19/1 x HT CIFC 832/1)	I	5,6,7,9
24	CIFC H147-1	S 353 4/5 (CIFC 34/13) x S 4 Agaro (CIFC 110/5)	T	2,3,4,5

* Dados extraídos de Bettencourt, 1981.

2.2. Extração de DNA

Folhas jovens de cada um desses cafeeiros foram coletadas, armazenadas em *ultrafreezer* -80^{o}C e liofilizadas. O DNA foi extraído seguindo-se o protocolo de Diniz *et al*. (2005). Após a extração, o DNA foi quantificado em espectrofotômetro e armazenado a 4^{o}C. Para a amplificação, o DNA foi diluído em tampão TE (Tris HCl 10 mM, EDTA 1 mM, pH 8,0) para uma concentração final de 10 ng/µl.

2.3. Amplificação de DNA

A reação de PCR foi realizada em 20,0µL de solução contendo 30ng de DNA, 0,8 unidades de *Taq* DNA polimerase, tampão 1x, 1mM de MgCl2, 60 µM de cada dNTP e 0,7 µM de cada *primer*. Para a amplificação, utilizou-se o procedimento *touchdown* PCR. Este procedimento consistiu em uma etapa de desnaturação inicial de 94^{o}C, por 3 minutos, seguido de cinco ciclos com uma etapa de desnaturação a 94^{o}C por 30 segundos, uma etapa de anelamento por 20 segundos e extensão a 72^{o}C por 40 segundos. A temperatura de anelamento foi de 65^{o}C a 60^{o}C, reduzindo 1^{o}C a cada ciclo. Após os cinco ciclos, procederam-se mais 30 ciclos com desnaturação a 94^{o}C por 60 segundos, anelamento de 60^{o}C por 20 segundos e extensão de 72^{o}C por 40 segundos. Após os ciclos, uma extensão final a 72ºC por 7 minutos.

Os produtos das reações de amplificação foram submetidos a eletroforese em géis de agarose 1,2%, os quais foram corados com brometo de etídio, visualizados em transiluminador sob luz UV e fotodocumentados. Os *primers* que apresentaram um perfil de bandas monomórficas foram submetidos à reação de purificação e sequenciamento para posterior identificação dos SNPs.

2.4. Purificação

Os produtos de PCR foram purificados com o uso da reação enzimática de Exonuclease I e Fosfatase Alcalina (ExoSAP-IT® - USB) na proporção de 5,0 µL de PCR para 2,0 µL de ExoSAP-IT. A reação foi incubada por 15 minutos a 37ºC, seguida da desativação por 15 minutos a 80ºC, como recomendado pelo fabricante. Posteriormente, o conteúdo de DNA dos produtos de PCR foi quantificado e encaminhado para o sequenciamento.

2.5. Sequenciamento

Os produtos de amplificação foram sequenciados com o "ABI Prism® BigDye® Terminator v3.1 Cycle Sequencing Kit" (Applied Biosystems), no analisador genético ABI 3130 XL (Applied Biosystems). As reações de sequenciamento foram baseadas na técnica de terminação de cadeia (Sanger *et al*. 1977), utilizando dideoxinucleotídeos (ddNTPs) marcados com fluorescência. O protocolo utilizado foi o recomendado pelo fabricante. Os *primers* utilizados para o sequenciamento foram os mesmos das respectivas reações de PCR, nas direções *forward* e *reverse*.

2.6. Análise das sequências

Para analisar as sequências obtidas, foram utilizados os programas Sequencher® v. 4.10 (visualização de cromatogramas, *trimming*, clusterização e edição), ORF Finder (identificação de ORFs), BLAST (alinhamentos locais) e ClustalW (alinhamentos globais múltiplos).

2.7. Sequências estudadas

Os 15 genes selecionados para o estudo dos SNPs foram os amplificados pelos *primers* CARF 001 (disease resistance-like protein), CARF 003 (receptor-like kinase), CARF 006 (disease resistance protein (CC-NBS-LRR class, putative), CARF 011 (disease resistance-like protein), CARF 019 (putative resistance protein), CARF 022 (putative protein kinase), CARF 024 (receptor kinase Lecrk), CARF 025 (serine/threonine-protein kinase-like protein), CARF 028 (membrane protein), CARF 036 (disease resistance-like protein), CARF 039 (putative disease resistance gene homolog), CARF 040 (chitinase), CARF 050 (non-race specific disease resistance protein [NDR1]), CARF 054 (chalcone synthase) e CARF 055 (chalcone synthase) (Tabela 2).

Tabela 2 – Anotação dos genes utilizados para a busca de SNPs e as sequências dos *primers* que os amplificam.

Primer	BLAST	*Score*	*e-value*	*Forward*	*Reverse*
CARF 001	gi|24459841|emb|CAC82597.1| disease resistance-like protein [*Coffea arabica*]	259	2.00E-67	CAAGAAACAATGGCTGAGG	CAATAGAGTCGGTCGGTCTG
CARF 003	gi|42602161|gb|AAS21681.1| receptor-like kinase [*Arabidopsis thaliana*]	102	1.00E-20	CGTCCCACGAGAAGATGATAC	TGAGTTGCCGAAGAAGTTG
CARF 006	gi|30689664|ref|NP_195056.2| disease resistance protein (CC-NBS-LRR class), putative [*Arabidopsis thaliana*]	652	0.0	GTCCCTATTCTGCCACTTTG	CAGCCCTTCTTCTCTTCTTG
CARF 011	gi|24459849|emb|CAC82600.1| disease resistance-like protein [*Coffea canephora*]	164	2.00E-39	AGAAGATGCTCAACCCAGAC	CCCATACCAACCACTGAAAC
CARF 019	gi|16974114|emb|CAC95155.1| putative resistance protein [*Solanum lycopersicon*]	253	2.00E-66	ATCCACTGCCACTCTCATC	ACCACGGCTCATTGTAAGTC
CARF 022	gi|18855060|gb|AAL79752.1| putative protein kinase [*Oryza sativa*]	126	2.00E-29	ATTGCTGGGCTGTTCTACAC	ATTCCCTCTCTCCTTCGTC
CARF 024	gi|31324528|gb|AAP47579.1| receptor kinase Lecrk [*Gossypium hirsutum*]	257	2.00E-67	TGCCTTCTTTCCTAATCCTG	TATGCTTGGCTGTTCCATC
CARF 025	gi|11358841|pir| serine/threonine-protein kinase-like protein [*Arabidopsis thaliana*]	115	1.00E-11	TGCCATTTCCTGAGTGTG	TGGGAGTGATGATTTGACTG
CARF 028	gi|15231381|ref|NP_187364.1| membrane protein [*Arabidopsis thaliana*]	141	2.00E-32	CTAACCACCACCCTTCAATC	AAACACCACCATCCACAAC
CARF 036	gi|24459853|emb|CAC82602.1| disease resistance-like protein [*Coffea arabica*]	142	5.00E-33	CCGTCTGGTTTCAATCGTC	ATCTCCTGGCAATCTCTCTG
CARF 039	gi|27817976|dbj|BAC55740.1| putative disease resistance gene homolog [*Oryza sativa*]	142	7.00E-33	AAAGAGCAGGACTTCACGAC	CTCCACCTAAGCAAAGACAAC
CARF 040	gi|3451147|emb|CAA09110.1| chitinase [*Hevea brasiliensis*]	266	4.00E-70	GGATACAGCACAGCAGAGAG	AGCAGCCAGGACTACATTTC
CARF 050	gi|15232308|ref|NP_188696.1| non-race specific disease resistance protein (NDR1) [*Arabidopsis thaliana*]	135	5.00E-31	GCCTGATGCCACTTTGTTC	AGCCTACACCTTTGTCCTTC
CARF 054	gi|1345787|sp|P48387| chalcone synthase [*Camellia sinensis*]	470	1.00E-131	GGCTGAGAACAACAAAGGTG	AAGAATGAACACGACACAGG
CARF 055	gi|12229619|sp|Q9ZRS4| chalcone synthase [*Catharanthus roseus*]	187	2.00E-46	GCCTGTGTCCTGTTCATTC	GGTCATTCAACCAAGCAAG

3. RESULTADOS E DISCUSSÃO

As análises dos 15 genes escolhidos para este estudo permitiram verificar que os quatro genes amplificados pelos *primers* CARF 001, CARF 006, CARF 036 e CARF 055, não apresentaram qualquer polimorfismo entre os 24 indivíduos testados (monomórficos). Outros cinco genes amplificados pelos *primers* CARF 003, CARF 011, CARF 19, CARF 022 e CARF 039 apresentaram polimorfismos apenas para o indivíduo 12, *Coffea liberica* var. *dewevrei*. Esta espécie é originária da África Central e filogeneticamente próxima a *C. liberica*, nativa da África Ocidental (Carvalho *et al*., 1990). Devido às semelhanças entre essas espécies, alguns autores (Lebrun, 1941; Charrier e Berthaud, 1985)

consideram *C. dewevrei* como uma variedade de *C. liberica*. Estudos quimiotaxonômicos, no entanto, indicam diferenças entre essas duas espécies, que, assim, devem ser mantidas separadamente (Lopes, 1984), tal como foi previamente considerado por Carvalho e Monaco (1967). A análise da sequência desse cafeeiro mostrou grande quantidade de SNPs, quando comparado com os demais indivíduos. No entanto, os polimorfismos encontrados provavelmente estão relacionados com as variações interespecíficas e não necessariamente associadas com as características conferidas por cada gene em questão.

Considerando apenas os indivíduos da espécie *C. arabica* e os acessos de Híbrido de Timor, seis genes amplificados pelos *primers* CARF 024, CARF 025, CARF 028, CARF 040, CARF 050 e CARF 054 apresentaram polimorfismos entre os indivíduos testados.

Foram detectados 71 SNPs (Tabela 3). Os SNPs foram analisados de acordo com a variação da base nucleotídica, em transição ou transversão. Dos 71 SNPs, 34 foram transições (47,89%) e 37 transversões (52,11%) (Tabela 4). A taxa de transições/transversões foi de 0,9189. Em gengibre essa taxa foi de 0,90 (Chandrasekar *et al*., 2009). Embora as variações de transição ocorram numa frequência mais alta que as transversões em genomas eucariotos (Li e Graur, 1991), possivelmente como resultado de mecanismos moleculares pelos quais são geradas, existem evidencias que isso não é universal (Keller *et al*., 2007). Moruyama e Powell (1996) detectaram 54% de transversões em *Drosophila*, valor bem parecido com o encontrado em soja (52%) por Nasu *et al*. (2002). Em outro trabalho com cafeeiros, foi encontrado 80% de mutações causadas por transversões e 20% decorrentes de transições (Zarate *et al*., 2010). Apesar de o presente trabalho estar de acordo com os resultados encontrados por Zarate *et al*. (2010), ainda não se sabe ao certo se em cafeeiros as transversões são mais frequentes que as transições.

Tabela 3 – SNPs polimórficos obtidos a partir do sequenciamento de genótipos de cafeeiro, suas posições e as modificações resultantes.

SNP	*Primer*	Indivíduo	Posição	Consenso	Modificação	Aminoácido	Modificação
01	CARF 024	20	196	T/A	A/T	(S - Ser) **T**CC	(T- Thr) **A**CC
02*	CARF 024	20	211	G/C	C/G	NA (A - Ala) GC**G**	NA (A - Ala) GC**C**
03	CARF 025	14	181	A/T	C/G	(T- Thr) **A**CA	(P - Pro) **C**CC
04	CARF 025	14	183	A/T	C/G	(T- Thr) AC**A**	(P - Pro) CC**C**
05*	CARF 025	15	183	A/T	C/G	NA (T- Thr) AC**A**	NA (T- Thr) AC**C**
06	CARF 025	18	181	A/T	C/G	(T- Thr) **A**CA	(P - Pro) **C**CA
07*	CARF 025	18	183	A/T	C/G	NA (T- Thr) AC**A**	NA (T- Thr) AC**C**
08*	CARF 025	19	183	A/T	C/G	NA (T- Thr) AC**A**	NA (T- Thr) AC**C**
09*	CARF 025	21	183	A/T	C/G	NA (T- Thr) AC**A**	NA (T- Thr) AC**C**
10	CARF 028	01	217	G/C	A/T	(V - Val) **G**TT	(I - Ile) **A**TT
11	CARF 028	01	488	G/C	C/G	(R - Arg) A**G**A	(T - Thr) A**C**A
12	CARF 028	02	217	G/C	A/T	(V - Val) **G**TT	(I - Ile) **A**TT
13	CARF 028	04	217	G/C	A/T	(V - Val) **G**TT	(I - Ile) **A**TT
14	CARF 028	04	488	G/C	C/G	(R - Arg) A**G**A	(T - Thr) A**C**A
15	CARF 028	06	217	G/C	A/T	(V - Val) **G**TT	(I - Ile) **A**TT
16	CARF 028	07	217	G/C	A/T	(V - Val) **G**TT	(I - Ile) **A**TT
17	CARF 028	07	378	G/C	T/A	(L - Leu) TT**G**	(F - Phe) TT**T**
18	CARF 028	07	488	G/C	C/G	(R - Arg) A**G**A	(T - Thr) A**C**A
19	CARF 028	09	217	G/C	A/T	(V - Val) **G**TT	(I - Ile) **A**TT
20	CARF 028	09	378	G/C	T/A	(L - Leu) TT**G**	(F - Phe) TT**T**
21	CARF 028	09	488	G/C	C/G	(R - Arg) A**G**A	(T - Thr) A**C**A
22	CARF 028	10	217	G/C	A/T	(V - Val) **G**TT	(I - Ile) **A**TT
23	CARF 028	11	217	G/C	A/T	(V - Val) **G**TT	(I - Ile) **A**TT
24	CARF 028	13	318	G/C	T/A	(L - Leu) TT**G**	(F - Phe) TT**T**
25	CARF 028	13	378	G/C	T/A	(L - Leu) TT**G**	(F - Phe) TT**T**
26	CARF 028	14	318	G/C	T/A	(L - Leu) TT**G**	(F - Phe) TT**T**
27	CARF 028	15	217	G/C	A/T	(V - Val) **G**TT	(I - Ile) **A**TT
28	CARF 028	15	488	G/C	C/G	(R - Arg) A**G**A	(T - Thr) A**C**A
29	CARF 028	16	217	G/C	A/T	(V - Val) **G**TT	(I - Ile) **A**TT
30	CARF 028	16	378	G/C	T/A	(L - Leu) TT**G**	(F - Phe) TT**T**
31	CARF 028	16	488	G/C	C/G	(R - Arg) A**G**A	(T - Thr) A**C**A
32	CARF 028	17	318	G/C	T/A	(L - Leu) TT**G**	(F - Phe) TT**T**
33	CARF 028	17	378	G/C	T/A	(L - Leu) TT**G**	(F - Phe) TT**T**
34	CARF 028	18	318	G/C	T/A	(L - Leu) TT**G**	(F - Phe) TT**T**
35	CARF 028	18	378	G/C	T/A	(L - Leu) TT**G**	(F - Phe) TT**T**
36	CARF 028	19	318	G/C	T/A	(L - Leu) TT**G**	(F - Phe) TT**T**
37	CARF 028	19	378	G/C	T/A	(L - Leu) TT**G**	(F - Phe) TT**T**
38	CARF 028	20	488	G/C	C/G	(R - Arg) A**G**A	(T - Thr) A**C**A
39	CARF 028	21	318	G/C	T/A	(L - Leu) TT**G**	(F - Phe) TT**T**
40	CARF 028	21	378	G/C	T/A	(L - Leu) TT**G**	(F - Phe) TT**T**
41	CARF 028	22	318	G/C	T/A	(L - Leu) TT**G**	(F - Phe) TT**T**
42	CARF 028	23	318	G/C	T/A	(L - Leu) TT**G**	(F - Phe) TT**T**
43	CARF 028	23	378	G/C	T/A	(L - Leu) TT**G**	(F - Phe) TT**T**
44	CARF 028	24	217	G/C	A/T	(V - Val) **G**TT	(I - Ile) **A**TT
45	CARF 028	24	488	G/C	C/G	(R - Arg) A**G**A	(T - Thr) A**C**A
46	CARF 040	13	140	T/A	C/G	(N - Asn) A**A**T	(S - Ser) A**G**T
47	CARF 040	14	117	T/A	C/G	(S - Ser) **A**GC	(G - Gly) **G**GC
48	CARF 040	14	140	T/A	C/G	(N - Asn) A**A**T	(S - Ser) A**G**T
49*	CARF 040	14	271	A/T	G/C	NA (N - Asn) AA**T**	NA (N - Asn)
50	CARF 040	15	117	T/A	C/G	(S - Ser) **A**GC	(G - Gly) **G**GC
51	CARF 040	15	140	T/A	C/G	(N - Asn) A**A**T	(S - Ser) A**G**T
52	CARF 040	16	106	C/G	G/C	(K - Lys) AA**G**	(N - Asn) AA**C**
53	CARF 040	16	117	T/A	C/G	(S - Ser) **A**GC	(G - Gly) **G**GC
54	CARF 040	16	140	T/A	C/G	(N - Asn) A**A**T	(S - Ser) A**G**T
55	CARF 040	16	288	C/G	T/A	(G - Gly) **G**GT	(S - Ser) **A**GT
56	CARF 040	18	117	T/A	C/G	(S - Ser) **A**GC	(G - Gly) **G**GC
57	CARF 040	18	140	T/A	C/G	(N - Asn) A**A**T	(S - Ser) A**G**T
58	CARF 040	19	117	T/A	C/G	(S - Ser) **A**GC	(G - Gly) **G**GC
59	CARF 040	19	140	T/A	C/G	(N - Asn) A**A**T	(S - Ser) A**G**T
60	CARF 040	21	117	T/A	C/G	(S - Ser) **A**GC	(G - Gly) **G**GC
61	CARF 040	21	140	T/A	C/G	(N - Asn) A**A**T	(S - Ser) A**G**T
62*	CARF 040	21	271	A/T	G/C	NA (N - Asn) AA**T**	NA (N - Asn)
63	CARF 040	21	288	C/G	T/A	(G - Gly) **G**GT	(S - Ser) **A**GT
64	CARF 040	22	117	T/A	C/G	(S - Ser) **A**GC	(G - Gly) **G**GC
65*	CARF 050	16	58	A/T	G/C	NA (E - Glu) GA**A**	NA (E - Glu) GA**G**
66	CARF 050	16	155	T/A	C/G	(F - Phe) **T**TT	(L - Leu) **C**TT
67*	CARF 054	16	178	T/A	C/G	NA (L - Leu) **T**TG	NA (L - Leu) **C**TG
68	CARF 054	20	354	T/A	A/T	(D - Asp) GA**T**	(E - Glu) GA**A**
69	CARF 054	20	355	T/A	C/G	(W - Trp) **T**GG	(R - Arg) **C**GG
70*	CARF 054	22	178	T/A	C/G	NA (L - Leu) **T**TG	NA (L - Leu) **C**TG
71*	CARF 054	24	231	A/T	T/A	NA (A - Ala) GC**A**	NA (A - Ala) GC**T**

* mutações sinônimas

Tabela 4 – Número e porcentagem de transições e transversões detectadas na análise de SNPs.

Transições	**Ocorrências**	**%**
A-G	3	4,22
G-A	11	15,50
C-T	2	2,81
T-C	18	25,35
Sub-Total	34	47,89
Transversões	**Ocorrências**	**%**
C-A	0	0
C-G	1	1,40
T-A	2	2,81
T-G	0	0
A-C	7	9,86
G-C	9	12,67
A-T	1	1,40
G-T	17	23,94
Sub-Total	37	52,11
TOTAL	**71**	**100,00**

Os SNPs também foram classificados considerando a sua posição no códon e o seu efeito sobre o aminoácido codificado, em mutações sinônimas ou mutações não sinônimas. Das 71 substituições, 27 ocorreram na primeira posição do códon (38,02%), 15 na segunda posição (21,12%) e 29 na terceira (40,84%) (Figura 1).

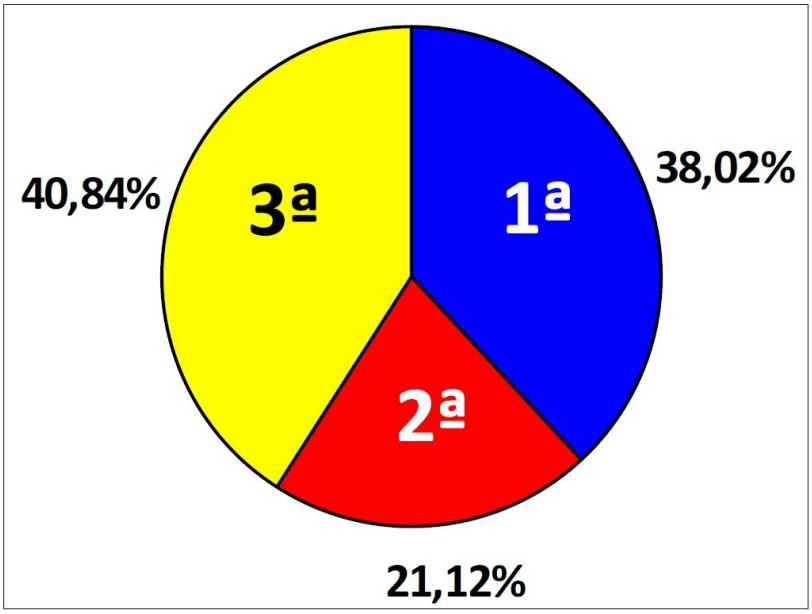


Figura 1 – Porcentagem de modificações de bases na primeira, segunda e terceira posição do códon.

É sabido que o pareamento do anticódon com a a terceira base no códon é menos rígido do que com as duas primeiras. Uma vez que a degenerância no código genético envolve, na maioria dos casos, somente a terceira base do

códon, o pareamento oscilante (*wobble base pairing*) explica as modificações sinônimas encontradas em vários organismos. Por outro lado, é possível relacionar a alta incidência de modificações na primeira e segunda base do códon com a elevada frequência de substituições não sinônimas encontradas neste trabalho.

Foram detectadas 11 (15,49%) substituições sinônimas e 60 (84,51%) substituições não sinônimas (Figura 2). A relação de substituições sinônimas/não-sinônimas foi de 0,1833. A frequência de substituições não sinônimas foi bastante elevada, se comparada com outras espécies. A taxa de modificações sinônimas para não sinônimas detectada em soja foi de 2,6 (Zhu *et al*., 2003) e em milho foi 4.8 (Tenaillon *et al*, 2001), bem abaixo da taxa de 8,7 encontrada em *D. melanogaster* (Moriyama e Powell, 1996). Em *Arabidopsis*, Oslen *et al*. (2002) encontrou taxas variando de 0,5 a 9,5 (média de 2,9) em seis genes analisados.

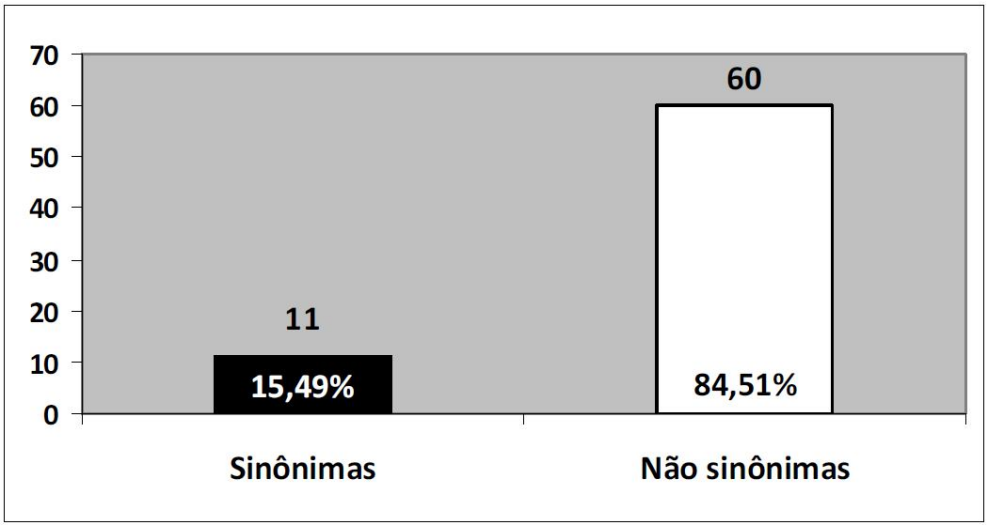


Figura 2 – Número e porcentagem de mutações sinônimas e não sinônimas detectadas.

Um alto nível de mutações não sinônimas, como o que foi encontrado neste trabalho, pode ser reflexo de seleção positiva (Meyerson e Sawyer, 2011). Um exemplo é o cenário antagonista da coevolução patógeno-hospedeiro, onde os genes do hospedeiro estão engajados numa "corrida armamentista" evolucionária e sob forte pressão de seleção para mudarem e se adaptarem. Em função disso, eles evoluem numa taxa mais rápida que outros genes (Kosiol *et al*., 2008). Genes de defesa estão entre os que evoluem com maior rapidez em mamíferos, adquirindo altos números de substituições não sinônimas (Gibbs *et*

al., 2007; Mikkelsen *et al*., 2005). Este cenário pode ser aplicado ao presente trabalho, dado o alto nível de modificações não sinônimas detectado e ao fato que os genes analisados atuam, de alguma forma, no processo de defesa do cafeeiro contra patógenos.

Considerando a análise da sequência dos 15 genes nos 23 cafeeiros da espécie *C. arabica* e Híbrido de Timor, foi observada uma frequência de 0,1029 SNPs por amplicon. Nos 119.061 pb analisados detectou-se uma frequência extremamente baixa de 1 SNP a cada 1.676,91pb, ou de 0,0596 SNPs por 100 pb. Salmaso *et al*. (2004) encontrou 1 SNP por 116 pb em cultivares de *Vittis vinifetra*. Em milho foi encontrado 1 SNP por 60-120 pb (Ching *et al*., 2002). A frequência média encontrada no arroz é de 1 SNP a cada 89 pb (Nasu *et al*., 2002) e no gengibre observou-se uma frequência de 0,84 SNPs por 100pb (Chandrasekar *et al*., 2009). Em outro trabalho com *Coffea*, Vidal *et al*. (2010) também obtiveram frequências baixas, variando entre 0,0409 e 0,0249 SNPs por 100 pb.

Pode-se atribuir o baixo nível de polimorfismo observado neste trabalho, à baixa diversidade do genoma de *C. arabica*. Seu modo reprodutivo autógamo é um dos fatores limitantes para a sua diversidade genética. O nível de polinização cruzada nesta espécie fica abaixo de 10% (Carvalho *et al.*,1991; Anthony *et al.*,2001). Outro fator relevante para a baixa diversidade de *C. arabica* é o modo como esta espécie se originou. Análises de citogenética molecular indicaram que *Coffea arabica* é um anfidiplóide formado a partir da hibridização entre duas espécies bem próximas (i.e. *Coffea canephora* ou *Coffea congensis* e *Coffea eugenioides*) (Raina *et al*., 1998; Lashermes *et al*., 1999). Outras análises sugeriram uma recente especiação e uma baixa divergência entre os dois genomas constitutivos de *Coffea arabica* e aqueles das suas espécies progenitoras (Fernandez e Lashermes, 2002). Um terceiro fator que contribui para a baixa diversidade da espécie é o modo como a planta foi introduzida no país. Evidências históricas sugerem que a introdução foi feita a partir de sementes e mudas trazidas da Guiana Francesa, e que a população-base descendeu de poucas árvores (Carvalho, 1993).

Por outro lado, sabe-se que *C. canephora* é altamente polimórfico. No entanto, os acessos de Híbrido de Timor (genótipos resistentes), que são constituídos, em média, por 90% de *C. arabica* e 10% de *C. canephora* (Setotaw, 2009) também apresentaram baixo polimorfismos com SNPs avaliados neste trabalho. A diversidade limitada observada em *C. arabica* dificulta a identificação de genes/alelos que fornecem resistência a estresses bióticos e abióticos. Por

esta razão muito se tem investido na busca por variabilidade genética por meio de novas fontes desse gênero.

O gene com o maior número de polimorfismos detectados foi o gi|15231381|ref|NP_187364.1| membrane protein [*Arabidopsis thaliana*], amplificado pelo *primer* CARF 028, com 36 SNPs (Tabela 3).

A análise dos produtos preditos do gene amplificado pelo *primer* CARF 028 permitiu verificar que as mutações não-sinônimas identificadas na sequência de DNA não originam proteínas distintas. Entretanto, se consideramos o cenário de co-evolução patógeno-hospedeiro, a alta ocorrência de mudanças de aminoácidos (i. e. substituições não sinônimas) pode contribuir para a melhoria da eficiência dessas proteínas. Análises mais detalhadas de modelagem tridimensional de proteínas poderão mostrar as possíveis diferenças que as modificações não sinônimas podem acarretar para as interações dos aminoácidos entre si (i.e. na cadeia popipeptídica) e com o meio.

O *primer* CARF 024 amplifica um fragmento cujo produto corresponde a região de 63-156 da proteína gi|31324528| gb|AAP47579.1| receptor kinase Lecrk [*Gossypium hirsutum*] de 641 aminoácidos. Essa proteína possui os domínios conservados lectin legume LecrK Arcelin ConA [cd06899] e PKc [cd00180] (Figura 3). Entretanto a região amplificada pelo *primer* CARF 024 apresenta apenas o domínio [cd06899].

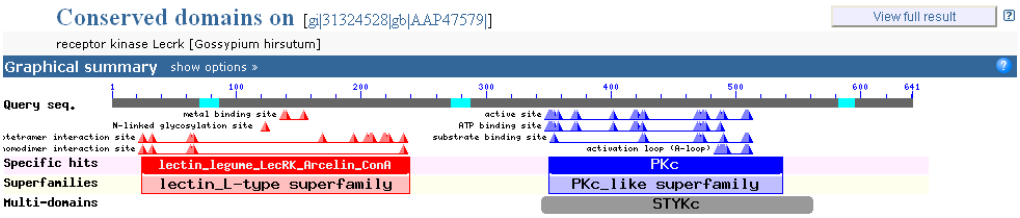


Figura 3 – Domínios conservados da proteína “receptor kinase LecRK”.

Os sítios importantes dos domínios conservados presentes na proteína predita amplificada pelo *primer* CARF 024 podem ser visualizados na Figura 4. É possível verificar também que a substituição S-T, causada por um SNP no indivíduo 20 não pertence a um sítio importante do domínio [cd06899].

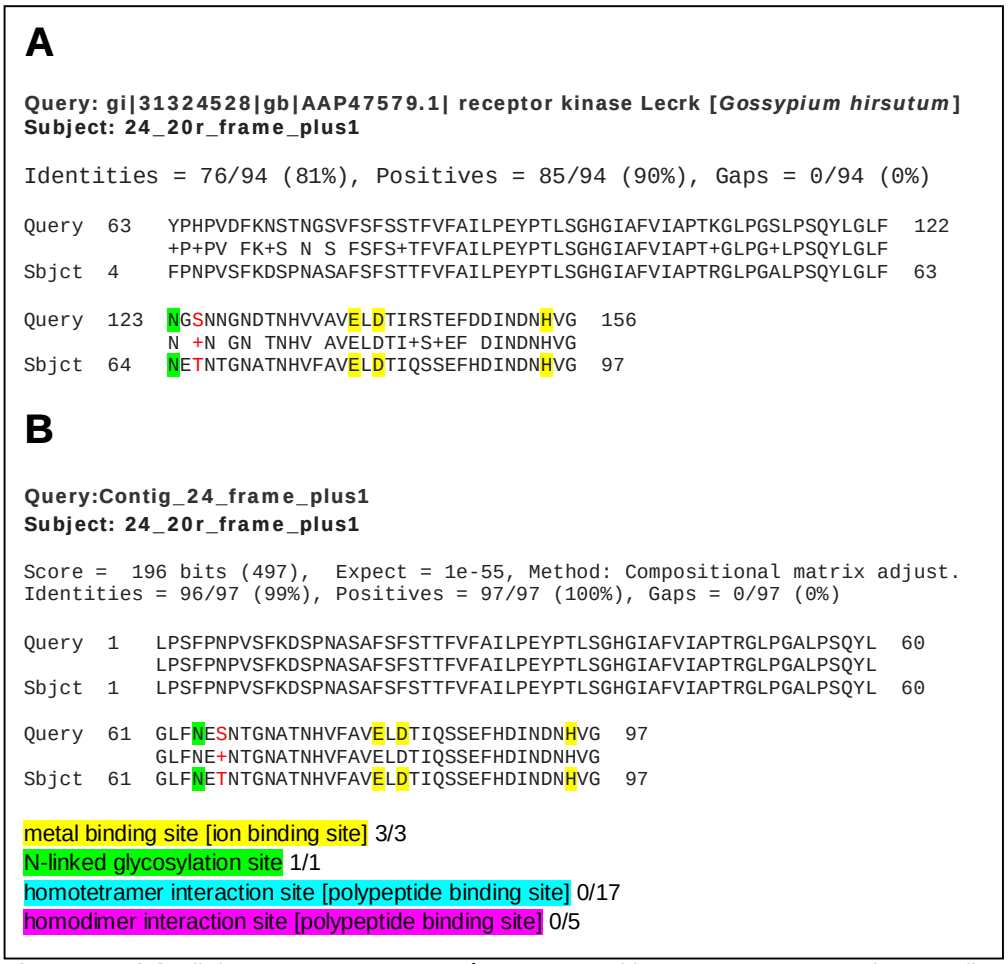


Figura 4 – (A) Alinhamento entre a proteína "receptor kinase LecrK" e o produto predito do gene amplificado pelo *primer* CARF 024, indivíduo 20. A substituição S-T (em vermelho) não atinge sítios funcionais da proteína (verde e amarelo). Os sítios funcionais "homotetramer interaction site" e "homodimer interaction site" (azul e ciano, respectivamente) da proteína "receptor kinase LecrK" não estão presentes no produto predito do gene amplificado pelo *primer* CARF 024, indivíduo 20. **(B)** Alinhamento entre os produtos preditos do *Contig* 024 e do gene amplificado pelo *primer* CARF 024 no indivíduo 20. A substituição T/A>A/T na posição 196 do DNA (Tabela 3) gera uma substituição S-T na proteína predita (em vermelho). Entretanto, essa substituição não atinge sítios funcionais da proteína (amarelo e verde). Os sítios funcionais "homotetramer interaction site" e "homodimer interaction site" (azul e ciano, respectivamente) da proteína "receptor kinase LecrK" não estão presentes nos produtos preditos do *Contig* 024 e do gene amplificado pelo *primer* CARF 024, indivíduo 20.

O *primer* CARF 040 amplifica a região 172-316 da proteína CAJ43737.1 GI:118200080 class III chitinase [*Coffea arabica*]. Essa proteína apresenta o domínio conservado GH18 hevamine XipI class III [cd02877] (Figura 5).

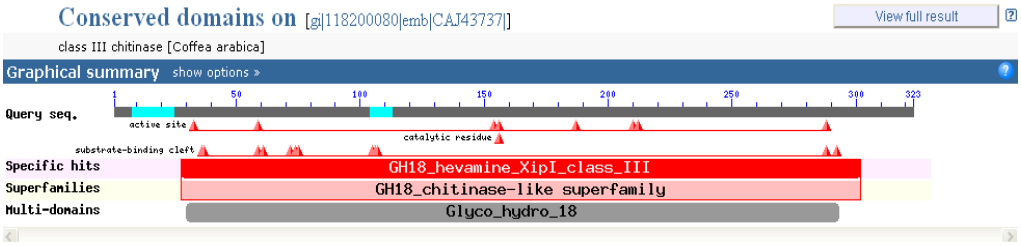


Figura 5 – Domínio conservado da proteína "class III chitinase"

Os sítios importantes dos domínios conservados presentes na proteína predita amplificada pelo *primer* CARF 040 podem ser visualizados na Figura 6. É possível verificar também que as substituições N-S, S-G e K-N, causadas por SNPs no indivíduo 16 não pertencem a um sítio importante do domínio [cd02877].

A

Query: gi|118200080|emb|CAJ43737|class III chitinase [*Coffea arabica*]
Subject: CARF_40_16R_frame-1
Score = 69.7 bits (169), Expect = 1e-17, Method: Compositional matrix adjust.
Identities = 36/57 (63%), Positives = 45/57 (79%), Gaps = 3/57 (5%)

```
Query  213  KIFLGLPAAPEAAGSG-YVPPDVLISRILPEIKKSPKYGGVMLWSKFYDDKNGYSSS  268
            K+FLGLPAA  AA SG Y+PP+VLIS+ILP ++  P YGGVMLWS++YD    YSS+
Sbjct  4    KLFLGLPAAQAAAPSGGYIPPEVLISQILPVVQGYPNYGGVMLWSRYYDQN--YSSA  58
```

B

Query: Contig40
Subject: CARF_40_16R_frame-1
Score = 111 bits (277), Expect = 2e-30, Method: Compositional matrix adjust.
Identities = 55/58 (95%), Positives = 56/58 (97%), Gaps = 0/58 (0%)

```
Query  70   GVNKLFLGLPAAQAAAPSGGYIPPEVLINQILPVVQSYPKYGGVMLWSRYYDQNYSSA  127
            GVNKLFLGLPAAQAAAPSGGYIPPEVLI+QILPVVQ YP YGGVMLWSRYYDQNYSSA
Sbjct  1    GVNKLFLGLPAAQAAAPSGGYIPPEVLISQILPVVQGYPNYGGVMLWSRYYDQNYSSA  58
```

active site [active site] 1/8
substrate-binding cleft [chemical binding site] 1/12
catalytic residue [active site] 0/1

Figura 6 – (A) Alinhamento entre a proteína "class III chitinase" e o produto predito do gene amplificado pelo *primer* CARF 040, indivíduo 16. As substituições N-S, S-G e K-N (em vermelho) não atigem sítios funcionais da proteína (verde e amarelo). O sítio funcional "catalytic residue" (em azul) da proteína "class III chitinase" não está presente no produto predito do gene amplificado pelo *primer* CARF 040, indivíduo 16. **(B)** Alinhamento entre os produtos preditos do *Contig* 040 e do gene amplificado pelo *primer* CARF 040 no indivíduo 16. As substituições T/A>C/G, T/A>C/G e C/G>G/C, nas posições 140, 117 e 106 do DNA (Tabela 3) geram substituições N-S, S-G e K-N na proteína predita (em vermelho). Entretanto, essas substituições não atingem sítios funcionais da proteína (amarelo e verde). O sítio funcional "catalytic residue" (em azul) da proteína "class III chitinase" não está presente nos produtos preditos do *Contig* 040 e do gene amplificado pelo *primer* CARF 040, indivíduo 16.

O *primer* CARF 025 amplifica uma região da proteína gi|11358841|pir|T47988 serine/threonine-protein kinase-like protein [*Arabidopsis thaliana*], mas não foram identificados domínios conservados para esta proteína até o momento. Desta forma, não é possível afirmar se os SNPs não-sinônimos encontrados nesse gene podem afetar algum sítio importante dessa proteína.

O *primer* CARF 028 amplifica a região 1-37 da proteína gi|15231381|ref|NP_187364.1| membrane protein [*Arabidopsis thaliana*]. Essa proteína possui os domínios EamA super family [cl01037] e RhaT [COG0697], porém nenhum dos dois está presente na região amplificada por este primer.

O *primer* CARF 050 amplifica a região 1-81 da proteína gi|15232308| ref|NP_188696.1| non-race specific disease resistance protein (NDR1) [*Arabidopsis thaliana*] mas não foram identificados domínios conservados para esta proteína até o momento. Desta forma, não é possível afirmar se os SNPs não-sinônimos encontrados nesse gene podem afetar algum sítio importante dessa proteína.

O *primer* CARF 054 amplifica a região 92-244 da proteína gi|1345787|sp|P48387| chalcone synthase [*Camellia sinensis*]. Esta proteína apresenta o domínio conservado CHS_like [cd00831] (Figura 7). Apesar de as sequências amplificadas conterem sítios importantes do domínio conservado em sua estrutura, nenhum deles coincidiu com a presença de SNP.

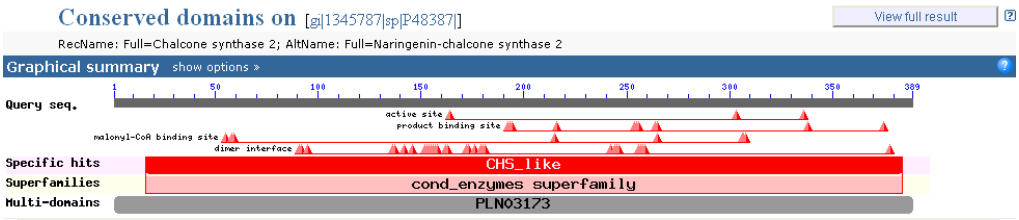
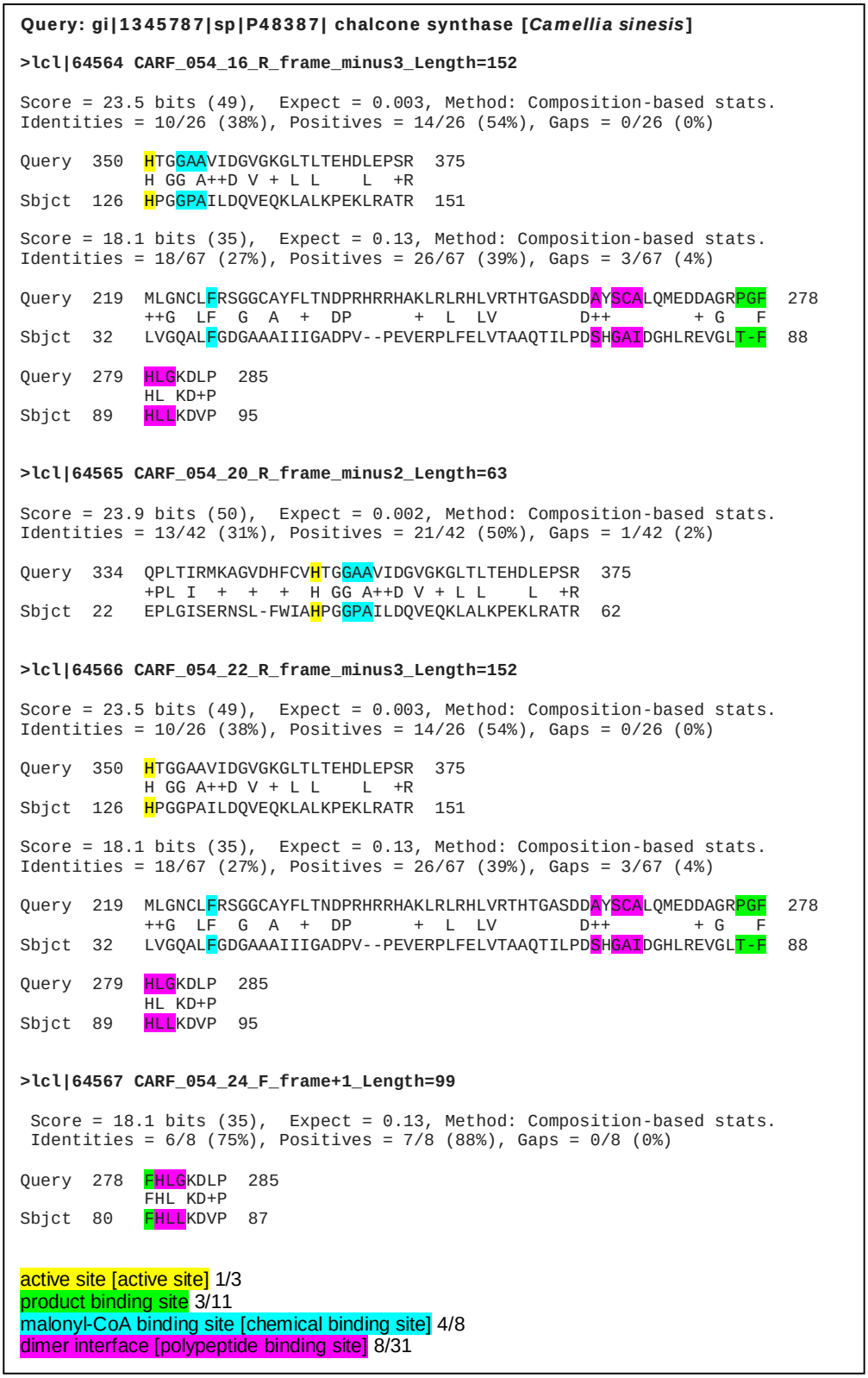


Figura 7 – Domínio conservado da proteína “chalcone synthase”

Os sítios importantes do domínio conservado CHS_like [cd00831] presentes na proteína predita amplificada pelo *primer* CARF 054 podem ser visualizados na Figura 8. Apesar de as sequências amplificadas conterem sítios importantes do domínio conservado em sua estrutura, nenhum deles coincidiu com a presença de SNP.



Query: gi|1345787|sp|P48387| chalcone synthase [*Camellia sinesis*]

>lcl|64564 CARF_054_16_R_frame_minus3_Length=152

Score = 23.5 bits (49), Expect = 0.003, Method: Composition-based stats.
Identities = 10/26 (38%), Positives = 14/26 (54%), Gaps = 0/26 (0%)

```
Query  350  HTGGAAVIDGVGKGLTLTEHDLEPSR  375
            H GG A++D V + L L    L  +R
Sbjct  126  HPGGPAILDQVEQKLALKPEKLRATR  151
```

Score = 18.1 bits (35), Expect = 0.13, Method: Composition-based stats.
Identities = 18/67 (27%), Positives = 26/67 (39%), Gaps = 3/67 (4%)

```
Query  219  MLGNCLFRSGGCAYFLTNDPRHRRHAKLRLRHLVRTHTGASDDAYSCALQMEDDAGRPGF  278
            ++G  LF  G  A  +  DP      +  L  LV        D++        + G   F
Sbjct  32   LVGQALFGDGAAAIIIGADPV--PEVERPLFELVTAAQTILPDSHGAIDGHLREVGLT-F  88

Query  279  HLGKDLP  285
            HL KD+P
Sbjct  89   HLLKDVP  95
```

>lcl|64565 CARF_054_20_R_frame_minus2_Length=63

Score = 23.9 bits (50), Expect = 0.002, Method: Composition-based stats.
Identities = 13/42 (31%), Positives = 21/42 (50%), Gaps = 1/42 (2%)

```
Query  334  QPLTIRMKAGVDHFCVHTGGAAVIDGVGKGLTLTEHDLEPSR  375
            +PL I  +  +  +  H GG A++D V + L L    L  +R
Sbjct  22   EPLGISERNSL-FWIAHPGGPAILDQVEQKLALKPEKLRATR  62
```

>lcl|64566 CARF_054_22_R_frame_minus3_Length=152

Score = 23.5 bits (49), Expect = 0.003, Method: Composition-based stats.
Identities = 10/26 (38%), Positives = 14/26 (54%), Gaps = 0/26 (0%)

```
Query  350  HTGGAAVIDGVGKGLTLTEHDLEPSR  375
            H GG A++D V + L L    L  +R
Sbjct  126  HPGGPAILDQVEQKLALKPEKLRATR  151
```

Score = 18.1 bits (35), Expect = 0.13, Method: Composition-based stats.
Identities = 18/67 (27%), Positives = 26/67 (39%), Gaps = 3/67 (4%)

```
Query  219  MLGNCLFRSGGCAYFLTNDPRHRRHAKLRLRHLVRTHTGASDDAYSCALQMEDDAGRPGF  278
            ++G  LF  G  A  +  DP      +  L  LV        D++        + G   F
Sbjct  32   LVGQALFGDGAAAIIIGADPV--PEVERPLFELVTAAQTILPDSHGAIDGHLREVGLT-F  88

Query  279  HLGKDLP  285
            HL KD+P
Sbjct  89   HLLKDVP  95
```

>lcl|64567 CARF_054_24_F_frame+1_Length=99

Score = 18.1 bits (35), Expect = 0.13, Method: Composition-based stats.
Identities = 6/8 (75%), Positives = 7/8 (88%), Gaps = 0/8 (0%)

```
Query  278  FHLGKDLP  285
            FHL KD+P
Sbjct  80   FHLLKDVP  87
```

active site [active site] 1/3
product binding site 3/11
malonyl-CoA binding site [chemical binding site] 4/8
dimer interface [polypeptide binding site] 8/31

Figura 8 – Alinhamento entre a proteína “chalcone synthase” e os produtos preditos dos genes amplificados pelo *primer* CARF 054, indivíduos 16, 20, 22 e 24. Os SNPs identificados nas sequências de DNA não estão presentes em sítios funcionais da proteína “chalcone synthase” (amarelo, verde, azul e ciano).

4. CONCLUSÕES

Neste trabalho foram detectados 71 SNPs, sendo que nenhum distinguiu indivíduos resistentes dos susceptíveis. Somente o gene “gi|15231381 |ref|NP_187364.1| membrane protein [*Arabidopsis thaliana*]”, amplificado pelo *primer* CARF 028, apresentou polimorfismo nos indivíduos susceptíveis. O baixo número de polimorfismos encontrados reflete a baixa diversidade do genoma de *C. arabica*. Mesmo nos acessos de Híbridos de Timor, constituídos de cerca de10% do genoma de *C. canephora*, altamente polimórfico, foram encontrados poucos polimorfismos. A análise dos produtos preditos permitiu verificar que as mutações não-sinônimas identificadas não estão presentes em sítios funcionais importantes nas proteínas analisadas. Entretanto, considerando o cenário da corrida armamentista evolucionária, a alta ocorrência de mudanças de aminoácidos (i. e. substituições não sinônimas) pode contribuir para a melhoria da eficiência dessas proteínas. Análises na estrutura tridimensional dos produtos preditos poderão revelar se as substituições identificadas neste trabalho causam algum impacto na geometria da proteína.

Os resultados encontrados no presente trabalho indicam que uso dos SNPs não deve ser a melhor estratégia para encontrar marcas polimórficas para esta espécie de diversidade tão reduzida.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Alvarenga SM (2007) **Caracterização de sequências expressas do genoma café potencialmente relacionadas com a resistência a doenças**. Dissertação. Genética e Melhoramento, Universidade Federal de Viçosa (UFV), Viçosa, 107p.

Anthony F, Bertrand B, Quiros O, Wilches A, Lashermes P, Berthaud J, Charrier A (2001) Genetic diversity of wild coffee (*Coffea arabica* L.) using molecular markers. **Euphytica**, 118:53-65.

Bent AF (1996) Plant disease resistance genes: function meets structure. **Plant Cell**, 8:1757–1771.

Bettencourt AJ (1981) **Melhoramento genético do cafeeiro**. Centro de Investigação das Ferrugens do Cafeeiro (CIFC), Lisboa, Portugal, 93p.

Bettencourt AJ, Noronha-Wagner M (1971), Genetic factors conditioning resistance of *Coffea arabica* L. to *Hemileia vastatrix* Berk and Br. **Agronomia Lusitana**, 31:285-292.

Bettencourt AJ, Noronha-Wagner M, Lopes M (1980), Factor genético que condiciona a resistência do clone 1343/269 (Híbrido de Timor) à *Hemileia vastatrix* Berk. and Br. **Brotéria Genética**, 1:53-58.

Carvalho A (1993) Histórico do desenvolvimento da cultura do café no Brasil. Instituto Agronômico de Campinas, Campinas, v.9, n.34, 7p. (Documento IAC).

Carvalho A, Fazuoli LC, Teixeira AA, Filho OG (1990) Aproveitamento do café excelsa em mistura com o café arábica. **Bragantia**, 48:335-343.

Carvalho A, Medina-Filho HP, Fazuoli LC, Guerreiro-Filho O, Lima MMA (1991) Aspectos genéticos do cafeeiro. **Revista Brasileira de Genética**, 14:135-183.

Carvalho A, Monaco LC (1967) Genetic relationships of selected *Coffea* species. **Ciência e Cultura**, 19:151-165.

Chandrasekar A, Riju A, Sithara K, Anoop S e Eapen SJ (2009) Identification of single nucleotide polymorphism in ginger using expressed sequence tags. **Bioinformation**, 4:119-122.

Charrier A, Berthaud J (1985) **Botanical classification of coffee**. In: Clifford MM e Willson KC, eds. Coffee: botany, biochemistry, and production of beans and beverage. Westport AVI Publishing, pp. 13-47.

Ching ADA, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Rafalski A (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. **BMC Genetics**, 3:19.

De Vienne D, Leonardi A, Damerval C, Zivy M (1999) Genetics of proteome variation for QTL characterization: application to drought-stress responses in maize. **Journal of Experimental Botany**, 50:303–309.

Diniz LEC, Sakiyama NS, Lashermes P, Caixeta ET, Oliveira ACB, Zambolim EM, Loureiro ME, Pereira AA, Zambolim L (2005) Analysis of AFLP markers associated to the Mex-1 resistance locus in Icatu progenies. **Crop Breeding and Applied Biotechnology**, 5:387-393.

Eskes AB (1989) **Resistance**. In: Kushalappa AC, Eskes AB (eds) Coffee rust: epidemiology, resistance and management. CRC Press, Boca Raton, pp 171–291

Fernandez D, Lashermes P (2002) Molecular tools for improving coffee (*Coffea arabica* L.) resistance to parasites. Molecular techniques. In: Join SM; Brar DS, Ahloowalia BS (eds) **Crop Improvement**. Kluwer Accdanc Publisher: Dordiecht, The Netherlands. p. 327-346.

Flor HH (1971) Current status of the gene- for- gene concept. **Annual Review of Phytopathology**, 9:275–296.

Gibbs RA *et al*. (2007) Evolutionary and biomedical insights from the rhesus macaque genome. **Science**, 316:222–234.

Hammond-Kosack KE, Jones DG (1996) Resistance gene dependent plant defense responses. **Plant Cell**, 8:1773–1791.

Hindorff LA, Junkins HA, Mehta JP, Manolio AT. **A catalog of published genome-wide association studies**. (http://www.genome.gov/page.cfm?pageid=26525384& clearquery=1#result_table).

Keller I, Bensasson D, Nichols RA (2007) Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes. **PLoS Genetics**, 3:e22.

Kosiol C. *et al*. (2008) Patterns of positive selection in six mammalian genomes. **PLoS Genetics**, 4:e1000144.

Lamb CJ, Lawton MA, Dron M, Dixon RA (1989) Signals and transduction mechanisms for activation of plant defenses against microbial attack. **Cell**, 56:215–224.

Lashermes P, Combes MC, Robert J, Trouslot P, D'hont A, Anthony F, Charrier A (1999) Molecular characterisation and origin of the *Coffea arabica* L. genome. **Molecular Genomics and Genetics**, 261:259-266.

Lebrun J (1941) Recherches morphologiques et systématiques sur /es ceféiers du Congo. Bruxelles, Instituí National pour l'Étude Agronomique du Congo Belga, 183p. (**Mémoire de la Section des Sciences Nature lies et Médicales**, coll. in - 8°, t XI).

Li WH, Graur D (1991) **Fundamentals of molecular evolution**. Sinauer Associates, Sunderland, MA, 284p.

Lopes CR (1984) **Estudos sobre as relações filogenéticas entre algumas espécies de *Coffea*.** Botucatu, UNESP, 1984. 171p. Dissertação (Livre-Docência).

Meyers BC, Dickerman AW, Michelmore RW, Sivaramakrishnan S, Sobral BW, Young ND (1999) Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. **Plant Journal**, 20:317–332.

Meyersonan NR, Sawyer SL (2011) Two-stepping throughtime: mammals and viruses. **Trends in Microbiology**,1–9.

Mikkelsen TS *et al*. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. **Nature**, 437:69–87.

Moriyama, E. N., and J. R. Powell, 1996 Intraspecific nuclear DNA variation in Drosophila. **Molecular Biology and Evolution,**13: 261–277.

Nasu S, Suzuki J, Ohta R, Hasegawa K, Yui R, Kitazawa N, Monna L e Minobe Y (2002) Search for and Analysis of Single Nucleotide Polymorphisms (SNPs) in Rice (*Oryza sativa*, *Oryza rufipogon*) and Establishment of SNP Markers. **DNA Research**, 9:163–171.

Noronha-Wagner M, Bettencourt AJ (1967), Genetic study of resistance of *Coffea* sp. to leaf rust. Identification and behaviour of four factors conditioning disease reaction in *Coffea arabica* to twelve physiologic races of *Hemileia vastatrix*. **Canadian Journal of Botany**, 45:2021-2031.

Olsen KM, Womack A, Garrett AR, Suddith JI e Purugganan MD (2002) Contrasting evolutionary forces in the *Arabidopsis thaliana* floral developmental pathway. **Genetics**, 160:1641–1650.

Pflieger S, Lefebvre V, Causse M (2001) The candidate gene approach in plant genetics: a review. **Molecular Breeding**, 7:275–291.

Raina SN, Mukai Y, Yamamoto M (1998) In situ hybridization identifies the diploid progenitor species of *Coffea arabica* (Rubiaceae). **Theoretical and Applied Genetics**, 97:1204-1209.

Salmaso M, Faes G, Segala C, Stefanini M, Salakhutdinov I, Zyprian E, Toepfer R, Grando MS, Velasco R (2004) Genome diversity and gene haplotypes in

the grapevine (*Vitis vinifera* L.), as revealed by single nucleotide polymorphisms. **Molecular Breeding**, 14:385–395.

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-termination inhibitors. **Proceedings of the National Academy of Sciences of the United States of America**, 74:5463-5467.

Setotaw, TA (2009) **Genetic diversity and genome introgression in coffee**. Tese. Genética e Melhoramento, Universidade Federal de Viçosa (UFV), Viçosa, 73p.

Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. mays L.). **Proceedings of the National Academy of Sciences of the United States of America,** 98:9161–9166.

Tsui C, Coleman LE, Griffith JL, Bennett AE, Goodson SG, Scott JD, Pittard SW, Devine SE (2003) Single nucleotide polymorphisms (SNPs) that map to gaps in the human SNP map. **Nucleic Acids Research**, 31:4910–4916.

Varzea VMP, Rodrigues CJ Jr, Silva MCML, Gouveia M, Marques DV, Guerra-Guimaraes L, Ribeiro A (2002) **Resistência do cafeeiro a *Hemileia vastatrix*.** In: Zambolim L (ed) O estado da arte de tecnologias na produção de café. Departamento de Fitopatologia. UFV, Viçosa, pp 297–320.

Vidal RO, Mondego JMC, Pot D, Ambrósio AB, Andrade AC, Pereira LFP, Colombo CA, Vieira LGE, Carazzolle MF, Pereira, GAG (2010) 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.

Zarate LA, Cristancho MA, Moncada P (2010) Strategies to develop polymorphic markers for *Coffea arabica* L. **Euphytica**, 173:243–253.

Zhu T, Salmeron, J (2007) High-definition genome profiling for genetic marker discovery. **Trends in Plant Science**, 12:196-202.

Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND, Cregan PB (2003) Single-Nucleotide Polymorphisms in Soybean. **Genetics**, 163:1123-1134.

CAPÍTULO 3

IDENTIFICAÇÃO DE CLONES BAC CONTENDO GENE DE RESISTÊNCIA DE CAFEEIRO A *Hemileia vastatrix*

1. INTRODUÇÃO

Doenças e pragas constantemente afetam a produção do café. Dentre elas destaca-se a ferrugem alaranjada, causada pelo fungo *Hemileia vastatrix*, a qual pode acarretar grandes prejuízos à cultura cafeeira, se medidas de controle não forem adotadas O estudo e a caracterização de fatores genéticos que conferem resistência a este fungo são importantes para ampliar os conhecimentos da interação planta-patógeno. O sequenciamento de genoma de plantas tem facilitado e acelerado a identificação de genes desejáveis, possibilitando a sua manipulação subseqüente por meio de técnicas de genética molecular. A biotecnologia do cafeeiro foi potencializada com o Projeto Brasileiro do Genoma Café.

O Projeto Brasileiro do Genoma Café teve início em 2002 e resultou num banco de dados de aproximadamente 200 mil ESTs (Vieira *et al.,* 2006). Partindo desses dados, Alvarenga (2007) realizou análises *in silico* a fim de encontrar sequências de genes relacionados com o mecanismo de resistência do cafeeiro a doenças. Por meio da mineração desses dados foi possível identificar 14.060 ESTs associadas a esta característica. Visando verificar o envolvimento destas sequências com a resistência do cafeeiro à ferrugem, Alvarenga (2007) obteve 40 pares de *primers* para amplificar algumas das sequências mineradas. Utilizando as condições de reação e amplificação otimizadas, os 40 *primers* foram testados em 12 genótipos resistentes e 12 susceptíveis a *H. vastatrix*. Foi identificado um marcador polimórfico**,** denominado CARF 005. Este *primer* amplifica uma região do DNA que corresponde a uma ORF (*Open Reading Frame* – Janela Aberta de Leitura) parcial de *Coffea arábica*, que codifica uma proteína de resistência a doenças. Esse marcador amplificou fragmento de DNA nos indivíduos resistentes e não amplificou nos suscetíveis.

Para estudar em detalhes o gene de resistência a doenças amplificado pelo CARF 005, é preciso obter a sua sequência completa. Genes de resistência a doenças foram clonados e caracterizados em plantas mono e dicotiledôneas (Hammond-Kosack e Jones, 1997). A literatura relata mais de 50 genes de resistência clonados em várias espécies de plantas (Wenkai *et al*., 2006). Uma das formas de conseguir isso é por meio da identificação desse fragmento no genoma de *C. arabica* e posterior sequenciamento das regiões adjacentes.

O Laboratório de Biotecnologia Vegetal do IAPAR (Instituto Agronômico do Paraná, Londrina) construiu uma biblioteca de BACs (*Bacterial Artificial Chromosome* – Cromossomo Artificial de Bactéria). Essa biblioteca possui

56.832 clones contendo fragmentos de DNA genômico de Híbrido de Timor 832/2 (Cação *et al*., 2007).

Assim, com a utilização do marcador CARF 005 e a biblioteca de BAC de *C. arabica*, o objetivo do presente trabalho foi realizar um *screening* desta biblioteca utilizando o marcador CARF 005 para identificar o(s) clone(s) contendo o fragmento do gene de resistência a doenças. Essa constitui a etapa inicial da clonagem e análise desse gene de resistência do cafeeiro a ferrugem.

2. MATERIAL E MÉTODOS

A biblioteca de BAC de Híbrido de Timor 832/2 foi disponibilizada pelo IAPAR (Cação *et al*. 2007). Uma cópia dessa biblioteca se encontra no Laboratório de Biotecnologia do Cafeeiro, UFV, sendo composta por um total de 56.832 clones. Esses clones estão armazenados em 148 placas, as quais possuem 24 colunas e 16 linhas, totalizando 384 poços. Cada placa foi virtualmente dividida em duas meias placas, contendo, cada uma, 12 colunas e 16 linhas. Desta forma, cada meia placa contém 192 clones. Essas meias placas foram numeradas de 1A a 148B. Posteriormente, foi realizada a extração do DNA dos clones de cada uma das 296 meias placas em conjunto (*pools*). Ou seja, cada tudo de *eppendorf* continha um *pool* de DNAs correspondentes aos clones de uma meia placa. A extração do DNA foi realizada utilizando o protocolo para recuperação de DNA de BACs-Mini-Prep (Diola, 2009). O DNA extraído foi quantificado e diluído para a concentração de 10 ηg.μL^{-1}. Após a extração e diluição, esse material foi devidamente armazenado em *ultrafreezer* -80^{o}C no Laboratório de Biotecnologia do Cafeeiro da UFV.

Para a identificação dos clones positivos, ou seja, identificação de clones de BAC que continham o fragmento de gene de resistência a doenças amplificado pelo *primer* CARF 005, realizou-se um *screening* da biblioteca BAC. O *screening* consistiu em amplificar as amostras de *pools* de DNA por meio de PCR utilizando o *primer* específico CARF 005 (Alvarenga, 2007). As etapas de amplificação compreenderam em *hot-start* de 94°C por 180 segundos, seguido de 5 ciclos de 94°C por 30 s, 65°C por 20 s e 72°C por 40 s e 30 ciclos de 94°C por 60 s, 60°C por 20 s, 72 °C por 40 s. Foi realizada uma extensão adicional de 72°C por 7 m. Para um volume final de 20μL foram utilizados 3μL de solução de DNA (10ηg.μL-1), 2 μL de tampão de reação de PCR 10x (Invitrogen), 0,4 μL de $MgCl_2$ (50mM) (Invitrogen), 0,6 μL de dNTP (2mM cada) (Invitrogen), 0,7 μL de cada *primer* (2 μM) e 0,16 μL de Taq Polimerase (5U/μL) (Invitrogen). Como

controles positivos foram utilizadas amostras de DNA provenientes de genótipos de *C. arabica* que comprovadamente são amplificados pelo *primer* CARF 005 (Alvarenga, 2007). Como controle negativo, utilizou-se água.

Todos os produtos de reação de PCR foram submetidos à eletroforese em gel de agarose a 1%, a 110 Volts, por aproximadamente 40 minutos. Após a eletroforese, o gel foi colocado em uma bandeja contendo solução de brometo de etídeo (0,50mg/L), sob agitação, por 5 a 10 minutos. Em sequência, o gel foi submetido à luz ultravioleta e fotografado com transiluminador Eagle Eye® II (Stratagene).

Na identificação do(s) clone(s) positivo(s), utilizou-se o método de decomposição de *pools*, consistindo em 3 etapas:

1º etapa – Identificação da(s) meia(s) placa(s) contendo clone(s) positivo(s)

2º etapa – Identificação da(s) coluna(s) positiva(s), da(s) meia(s) placa(s) positivas encontrada(s) na etapa anterior

3º etapa – Identificação dos clones positivos, utilizando as colunas positivas obtidas na etapa anterior.

Para repicagem das placas foi utilizado, em cada poço, 70 µL do meio de cultura composto por 13mM de KH_2PO_4, 36mM de K_2HPO_4, 1,7mM citrato de sódio, 6,8 mM de $(NH_4)_2SO_4$, 5% de glicerol, 1% de triptona, 0,5% de extrato de levedura e NaCl. Após a inoculação das BACs, submeteu-se a placa a uma temperatura de 37°C, em aproximadamente 150 rpm de agitação por 24 horas.

3. RESULTADOS E DISCUSSÃO

Após a análise, por PCR, dos 296 *pools* de DNA, que correspondem as meias placas da biblioteca de BAC, foram identificados 2 *pools* positivos, as meias placas 78B e 144B (Figura 1).

A meia placa 78B foi repicada e na segunda etapa de decomposição dos *pools* desta meia placa foram identificados 2 *pools* positivos (78B_14 e 78B_18) (Figura 2). Na terceira etapa de decomposição, dois clones positivos (78B_14.G e 78B_14.I) pertencentes a coluna 78B_14 foram identificados (Figura 3).

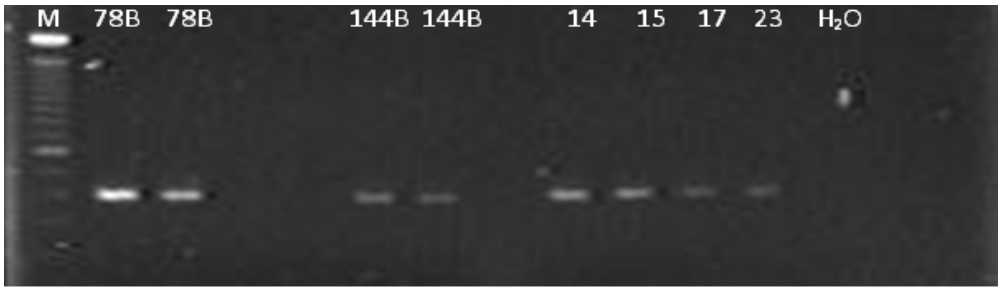


Figura 1 – Identificação de dois *pools* contendo clones positivos. M = Marcador de peso molecular 100pb. Meia placa 78B e 144B. Controles positivos: 14 (Híbrido de Timor UFV 443-3), 15 (Híbrido de Timor UFV 445-46), 17 (Híbrido de Timor CIFC 832/2) e 23 [CIFC H420-10 = Mundo Novo (CIFC 1535/33) x HW 26/14 (Caturra Vermelho CIFC 19/1 x HT CIFC 832/1)] H_2O = Controle negativo.

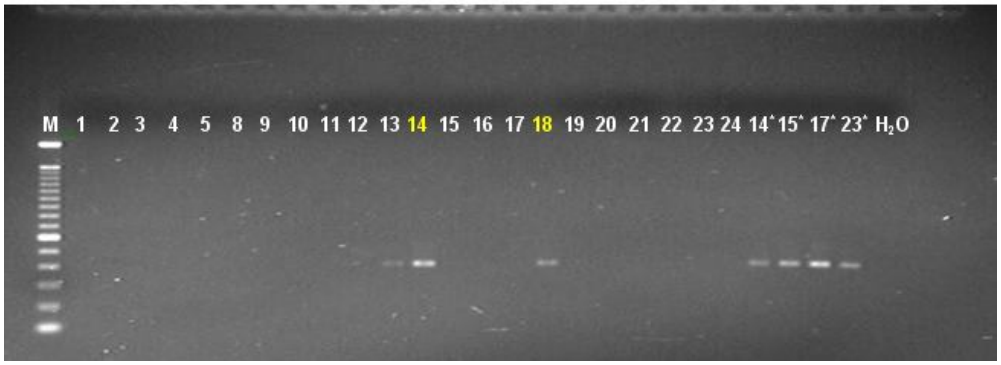


Figura 2 – Identificação das colunas contendo clones positivos da meia placa 78B. M = Marcador de peso molecular 100pb. 1-12 = Colunas da meia placa 77B. 13-24 = Colunas da meia placa 78B. 14* (Híbrido de Timor UFV 443-3), 15* (Híbrido de Timor UFV 445-46), 17* (Híbrido de Timor CIFC 832/2) e 23* [CIFC H420-10 = Mundo Novo (CIFC 1535/33) x HW 26/14 (Caturra Vermelho CIFC 19/1 x HT CIFC 832/1)] = Controles positivos. H_2O = Controle negativo.

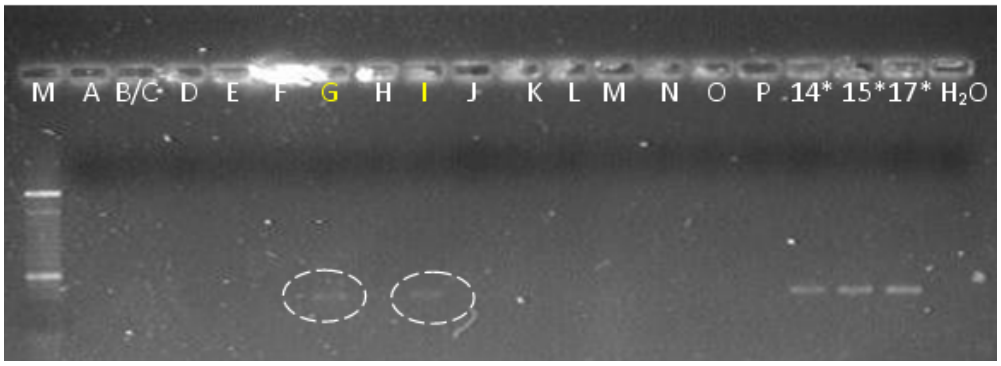


Figura 3 – Identificação dos clones positivos da coluna 14, da meia placa 78B. A-P = Linhas da meia placa 78B_14. 14* (Híbrido de Timor UFV 443-3), 15* (Híbrido de Timor UFV 445-46) e 17* (Híbrido de Timor CIFC 832/2) = Controles positivos. H2O = Controle negativo.

O resumo dos resultados obtidos nas etapas de decomposição dos *pools* está apresentado na Figura 4. Nessa figura é possível observar as etapas de

decomposição de *pools*. A 1ª etapa constituiu na identificação de meias placas positivas. A 2ª etapa, a identificação de colunas positivas utilizando meias placas positivas encontradas na 1ª etapa de decomposição. A 3ª etapa apresenta a identificação do(s) clone(s) positivo(s) utilizando colunas positivas encontradas. Nesta figura são representadas especificamente as etapas positivas da meia placa 78B, que permitiu identificar dois clones positivos.

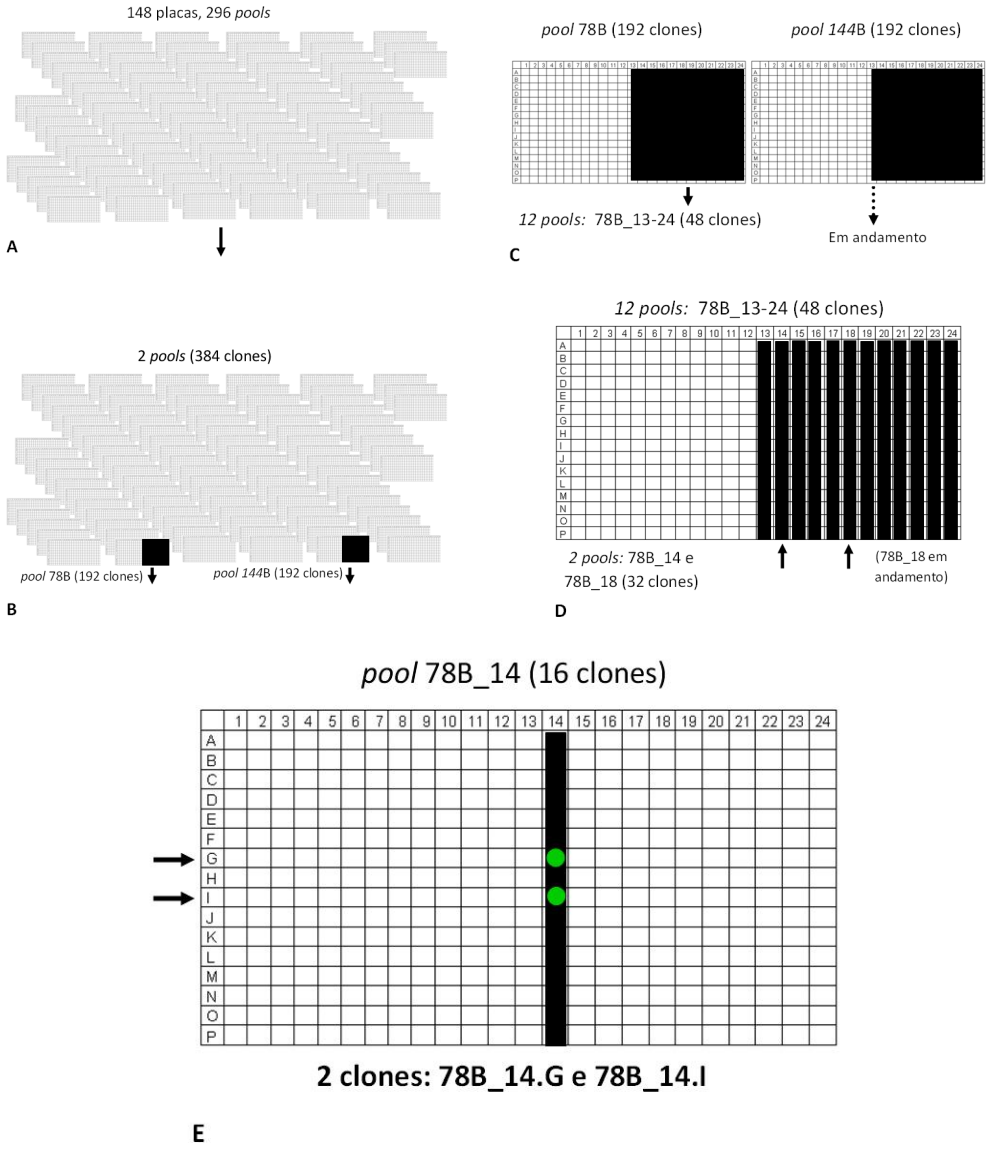


Figura 4 – Representação das etapas utilizadas para a identificação dos clones que possuem fragmento de gene de resistência amplificado pelo *primer* CARF 005. Partindo de 148 placas (**A**) foram identificadas duas meias placas (**B**), na primeira etapa do *screening* (**C** - 78B e 144B). Da meia placa 78B identificou-se na etapa seguinte (**D**) duas colunas positivas (78B_14 e 78B_18). Na última etapa do *screening* (**E**), partindo da coluna 78B_14, foram identificados dois clones positivos (78B_14.G e 78B_14.I).

Os dois clones identificados podem fornecer informações importantes sobre a estrutura dos genes de resistência em *Coffea*. Após o sequenciamento, será possível analisar toda a extensão do gene de resistência amplificado parcialmente pelo CARF 005. Desta forma, será possível saber o número de íntrons e éxons que compõem esse gene. Além disso, será possível conhecer a região promotora do gene. Os promotores representam elementos essenciais que podem trabalhar em conjunto com outras regiões regulatórias para direcionar o nível de transcrição de um gene. A partir da análise dessas regiões, será possível ampliar o conhecimento sobre a modulação da expressão desse gene em casos de interação cafeeiro-patógeno.

Após os estudos sobre a estrutura do gene, a sua clonagem poderá ser usada na obtenção de plantas transgênicas. A transgenia é uma das tecnologias que pode ser utilizada nos programas de melhoramento de plantas. Essa tecnologia oferece duas oportunidades aos melhoristas. Uma é a introdução de uma nova variação genética que não se encontra disponível no germoplasma do programa de melhoramento, e a outra é a criação de fenótipos desejados a partir de genes conhecidos (Zhong, 2001).

O potencial da tecnologia de transformação genética no melhoramento de culturas foi bem demonstrado na comercialização de variedades e híbridos com novas características transgênicas como resistência a doenças e a insetos (During, 1996; Jouanin *et al*.,1998) e tolerância a herbicidas (Tsaftaris, 1996).

Em café, vários trabalhos envolvendo clonagem gênica e transformação já foram realizados. Entre eles destacam-se o gene que codifica a subunidade menor da rubisco (Marraccini *et al*., 2003), os três genes envolvidos na síntese de cafeína (Uefuji *et al*., 2003) e, posteriormente, a obtenção de café descafeinado (Ogita *et al*., 2003).

4. CONCLUSÕES

Este trabalho constituiu a primeira etapa da clonagem do gene de cafeeiro que confere resistência a ferrugem. Partindo de uma biblioteca BAC contendo 56.832 clones distintos, foi possível identificar dois clones contendo fragmento de gene de resistência amplificado pelo marcador CARF 005. Com a identificação destes clones será possível, posteriormente, realizar o sequenciamento da extensão total do gene. A análise da sequência completa do gene permitirá identificar e caracterizar a sequência promotora e terminadora, bem como determinar a localização e quantidade de íntrons e éxons. Outra

informação relevante que poderá ser obtida é a estrutura organizacional dos genes de resistência em *Coffea*. Com o sequenciamento das regiões adjacentes ao gene de resistência amplificado pelo *primer* CARF 005, será possível verificar se os genes de resistência estão organizados em *clusters*, como é comumente relatado para outras espécies. Espera-se, dessa forma, contribuir com o aumento da base do conhecimento do funcionamento deste gene no processo de defesa do cafeeiro e, de uma maneira geral, melhorar o entendimento da interação café-ferrugem.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Alvarenga SM (2007) **Caracterização de sequências expressas do genoma café potencialmente relacionados com a resistência a doenças.** Viçosa: UFV, 2007. 107p. Dissertação (Mestrado em Génetica e Melhoramento)- Universidade Federal de Viçosa.

Cação SMB, Diniz LC, Silva NVE, Vinecky F, Carvalho A, Pereira LFP, Vieira LG (2007) **Construção de uma biblioteca genômica de cromossomo artificial de bactéria de *Coffea arabica***. In: V Simpósio de Pesquisa dos Cafés do Brasil, 2007, Águas de Lindóia. Anais do V Simpósio de Pesquisa dos Cafés do Brasil. Brasíia : Embrapa CD-ROM.

Diola V (2009) **Resistência à ferrugem do cafeeiro: mapeamento genético, físico e análise de expressão gênica em resposta a infecção de *H. vastatrix***. Tese de doutorado. Universidade Federal de Viçosa. 90 p.

During K (1996) Genetic engineering for resistance to bacteria in transgenic plants by introduction of foreign genes. **Molecular Breeding**, 2: 297–305.

Hammond-Kosack KE, Jones JDG (1997) Plant disease resistance genes. **Annual Review of Plant Physiology and Plant Molecular Biology***,* 48:575-607.

Jouanin L, Bonade-Bottino M, Girard C, Morrot G, Giband M (1998) Transgenic plants for insect resistance. **Plant Science**, 131: 1–11.

Marraccini P, Courjault C, Caillet V, Lausanne F, Lepage B, Rogers WJ, Tessereau S, Deshayes A (2003) Rubisco small subunit of *Coffea arabica*: cDNA sequence, gene cloning and promoter 35 analysis in transgenic tobacco plants. **Plant Physiology and Biochemistry**, 41:17–25.

Ogita S, Uefuji H, Yamaguchi Y, Koizumi N, Sano H (2003) Producing decaffeinated coffee plants. **Nature,** 423:823.

Tsaftaris A (1996(The development of herbicide-tolerant transgenic crops. **Field Crops Research**, 45:115–123.

Uefuji H, Ogita S, Yamaguchi Y, Koizumi N, Sano H (2003) Molecular Cloning and Functional Characterization of Three Distinct *N*methyltransferases Involved in the Caffeine Biosynthetic Pathway in Coffee Plants. **Plant Physiology**, 132:372–380.

Vieira LGE, Andrade AC, Colombo CA, Moraes AHA, Metha A, Oliveira AC, Labate CA, Marino CL, Monteiro-Vitorello CB, Monte DC, *et al.* (2006) Brazilian coffee genome project: An EST-based genomic resource. **Brazilian Journal of Plant Physiology**, 18:95-108.

Wenkai X, Mingliang X, Jiuren Z, Fengge W, Jiansheng L, Jingrui D (2006) Genome-wide isolation of resistance gene analogs in maize (*Zea mays* L.). **Theoretical and Applied Genetics**,113:63-72.

Zhong GY (2001) Genetic issues and pitfalls in transgenic plant breeding. **Euphytica**, 118: 137–144.

3. CONCLUSÕES GERAIS

Este trabalho constituiu-se de uma continuação dos estudos iniciais sobre a identificação de genes de resistência a doenças no Genoma Café, realizados por Alvarenga (2007).

As análises detalhadas sobre as quitinases no capítulo 1 permitiram ressaltar a grande importância desta enzima para a célula vegetal. Os resultados obtidos corroboraram com estudos realizados anteriormente, ratificando o envolvimento das quitinases na defesa da planta contra patógenos e o seu envolvimento em resposta de defesa da célula vegetal contra estresses.

O estudo com SNPs, no capítulo 2, mostrou baixo nível de polimorfismo em *C. arabica*, fato que já foi demonstrado em vários estudos moleculares. Além disso, os resultados encontrados neste trabalho indicam que uso dos SNPs não deve ser a melhor estratégia para encontrar marcas polimórficas nesta espécie.

O terceiro capítulo tratou da primeira etapa da clonagem do gene de cafeeiro que confere resistência a ferrugem. A identificação dos clones positivos foi resultado muito animador para a continuação do trabalho. Várias informações essenciais sobre a estrutura tanto do gene amplificado pelo *primer* CARF 005, como também dos genes de resistência a doenças em *Coffea* poderão ser obtidas nas próximas etapas do trabalho. Ressalta-se a relevância desses resultados, visto que, para o nosso conhecimento, nenhum gene de resistência de cafeeiro a doenças foi clonado até o momento.