POLIMORFISMOS NUCLEOTÍDICOS DE GENES ENVOLVIDOS NAS CARACTERÍSTICAS QUÍMICAS DO GRÃO DE CAFÉ. COMPLEMENTARIDADE DAS ESTRATÉGIAS *IN SILICO* E *IN VIVO*

Sérgio Dias Lannes[1]: sdlannes@terra.com.br, Sophie Bouchet[4], Lucia Pires Ferreira[1], Thierry Leroy[4], Suzana Tiemi Ivamoto[1], Pierre Marracini[3,4], Luiz Filipe Protasio Pereira[5], Luiz Gonzaga Vieira[1], David Pot[1,4]
[1]Instituto Agronômico do Paraná, Londrina, PR, [2]Institut National des Sciences Appliquées, Toulouse, France, [3]Embrapa Recursos Genéticos e Biotecnologia, Brasília, DF, [4]CIRAD, UMR DAP Montpellier, France. [5]Embrapa Café, Londrina, PR.

Resumo: A compreensão das bases genéticas da composição química do grão de café é indispensável para a gestão dos programas de melhoramento que têm como objetivo a qualidade da bebida. O desenvolvimento da genômica permite hoje a identificação de genes candidatos que são potencialmente envolvidos nessas características. Entretanto, a utilização destas novas ferramentas para o melhoramento depende da capacidade de identificar dentro de todos esses candidatos,, os que controlam a variabilidade das características entre genótipos. Essa identificação envolve o teste das relações entre os polimorfismos dos genes e a variabilidade fenotípica. A avaliação da diversidade nucleotídica pode ser feita de duas maneiras: usando as informações disponíveis nos bancos de dados EST (estratégia *in silico*) ou por seqüênciamento direto de genótipos de interesse (estratégia *in vivo*). O objetivo desse estudo foi avaliar o potencial da estratégia de analise de polimorfismos *in silico* para o café baseado nos bancos de dados disponíveis. Foram estudadas as vias da biossíntese da sacarose e dos diterpenos, compostos com efeitos na qualidade da bebida e na saúde humana, respectivamente. Essa busca permitiu a identificação de 1.1 polimorfismos para cada 100 bp para os 14 genes estudados. Uma avaliação da diversidade nucleotídica *in vivo* para alguns desses genes (via da biossíntese da sacarose) permitiu comparar essas duas estratégias. A estratégia *in silico* é complementar à estratégia *in vivo* permitindo uma avaliação geral dos níveis de polimorfismos dos genes numa larga escala em todo o genoma com um baixo custo.

Palavras Chaves: Diversidade nucleotídica, Análise *in silico*, Bancos ESTs, Sacarose, Diterpenos

NUCLEOTIDE POLYMORPHISMS OF GENES INVOLVED IN CHEMICAL CHARACTERISTICS OF COFFEE BEAN. COMPLEMENTARITY OF THE IN SILICO AND IN VIVO STRATEGIES

Abstract: The understanding of the genetic bases of coffee bean chemical composition is a requirement for breeding programs, which aim to improve coffee beverage quality. Nowadays, the development of the genomic toolkit allows the identification of candidate genes potentially controlling these traits. However, the direct utilization of these new tools for breeding relies on the ability to identify within this set of candidates the ones that are responsible for the variability observed between genotypes, *i.e.* the ones whose polymorphisms affect the variability of the traits of interest. Evaluation of the nucleotide diversity of the genes, which is a pre-requisite to test these associations, can be done in two ways: through the analysis of the EST libraries available (*in silico* strategy) or through direct sequencing of genotypes of interest (*in vivo* strategy). The purpose of this study was to evaluate the relevance of the *in silico* polymorphism discovery strategy in *Coffea* based on the EST libraries currently available. Genes for the biosynthetic pathways leading to sucrose and coffee specific diterpens (cafestol and kawheol) were analysed *in silico*. This strategy yielded the identification of 1.1 polymorphism per 100 bp in average for the 14 analysed genes, this result underlying the feasibility of this method for Coffee. Analysis of the nucleotide diversity *in vivo* for a few genes of the sucrose biosynthesis pathway allowed a comparison of the two strategies. The *in silico* discovery strategy is complementary to the *in vivo* one providing in a cost effective manner a first evaluation of nucleotide diversity at the whole genome level.

Key Words: Nucleotide diversity, *in silico* analysis, EST libraries, Sucrose, Diterpenes

Introdução

O café é a segunda maior commodity do mercado mundial, sendo o Brasil maior produtor, exportador e o segundo consumidor. Apesar da importância no mercado internacional, ainda são poucos os dados disponíveis sobre a composição química dos grãos de café. Ela afeta diretamente a qualidade da bebida e tem também alguns efeitos na saúde humana. Se o local de cultivo, métodos de colheita, armazenamento e modo de preparo da bebida influenciam a composição química, o fator genético (espécie, variedade) tem também uma alta importância. A qualidade da bebida do café é o principal fator de agregação de valor ao produto. Atualmente, o mercado valoriza cada vez mais os produtos com características específicas de aroma e de sabor, independente do local de origem. Além disso, há demanda dos consumidores sobre um maior conhecimento dos compostos que podem ter efeitos na saúde humana.

Nesse contexto, o estudo das bases genéticas da composição química do grão é uma prioridade, oferecendo bases para o melhoramento genético para a qualidade e para a síntese de compostos bioquímicos de interesse para a saúde. Recentemente, devido ao grande avanço obtido nas áreas de seqüênciamento de genomas, uma nova classe de marcadores baseados em polimorfismos de nucleotídeos (SNPs – Single Nucleotide Polymorphism) tem ganhado destaque. SNPs são

marcadores moleculares com capacidade de diferenciar indivíduos por meio de variações em apenas um nucleotídeo de seqüências de DNA que podem ou não codificar genes. A busca destes marcadores é possível por duas maneiras: o seqüênciamento direto de genes candidatos de interesse dentro de populações (estratégia *in vivo*) ou pela análise de dados de ESTs disponíveis em bancos de dados (estratégia *in silico*).

O objetivo desse estudo é apresentar uma análise exploratória de busca de polimorfismos para genes envolvidos na biossíntese da sacarose e de diterpenos específicos do café (cafestol e caveol), dois compostos presentes nos grãos de café importantes para a qualidade da bebida e para a saúde humana, respectivamente. Uma busca dos polimorfismos desses genes foi realizada dentro dos bancos EST disponíveis (estratégia *in silico*) e para o metabolismo da sacarose foi feita uma analise de diversidade nucleotídica *in vivo* usando varias espécies de *Coffea*. A eficiência e a complementaridade das duas estratégias foram discutidas.

Material e Métodos

Busca *in silico*

A busca *in silico* das seqüências que correspondem às enzimas de interesse, que compõem as rotas metabólicas da sacarose e dos diterpenos foi realizada em três bancos de ESTs diferentes. O maior banco, denominado Genoma Café (http://www.lge.ibi.unicamp.br/cafe/, Vieira et al., 2006), é composto de 37 bibliotecas baseadas em dois genótipos de *Coffea arabica (Ca)*, um de *Coffea canephora (Cc)* e um de *Coffea racemosa (Cr).* As diferentes bibliotecas correspondem a diferentes tipos celulares, diferentes tecidos e diferentes fases de desenvolvimento, com cerca de 150.000 seqüências no total.

Outro banco de ESTs utilizado foi o desenvolvido pela Cornell University/Nestlé (Lin et al., 2005) sendo composto por cinco bibliotecas de cDNA principalmente de frutos de *Cc*, gerando cerca de 47.000 ESTs (Lin 2005).

O terceiro banco de ESTs foi desenvolvido pelo IRD (Institut de Recherche pour le Développement), na França, utilizando a espécie *Cc.* Esse banco possui duas bibliotecas de ESTs, sendo uma de folhas jovens com 4.606 seqüências e uma outra incluindo diferentes estádios de desenvolvimento de frutos com 5.814 seqüências.

As enzimas estudadas para a rota da sacarose foram a sacarose sintase (SUS), a invertase de parede (CWI), a invertase de vacúolo (VI) e a sacarose fosfato sintase (SPS). As enzimas envolvidas na rota metabólica dos diterpenos foram copalil fosfato sintase (CPS), caurene sintase (KS), caurene oxidase, isopentenil difosfato sintase (IDS) e 1-deoxi-D-xilulose 5-fosfato reductoisomerase(DXR).

As ESTs foram encontradas a partir das seqüências protéicas disponíveis em outros gêneros usando a ferramenta tblastn. Quando possível, os cromatogramas foram usados para a busca *in silico* dos polimorfismos (Genoma Café e Cornell/Nestlé). O alinhamento das seqüências, a busca dos polimorfismos e a definição dos tipos dos polimorfismos foram feitos com o software Codon Code Aligner (www.codoncode.com/) junto com as ferramentas disponíveis no site do NCBI (http://www.ncbi.nlm.nih.gov/). Além da avaliação da qualidade dos cromatogramas, só foram considerados os polimorfismos para aqueles alelos em freqüência maior que duas vezes, isso para evitar os erros de seqüênciamento.

Busca *in vitro* de genes envolvidos na biossíntese de sacarose

Para análise *in vivo* de alguns genes envolvidos na via da biossíntese da sacarose foi feita amplificação de seqüências genômicas de diferentes genótipos de *Cc* e de 14 outras espécies de *Coffea.* O desenho dos nucleotídeos utilizados foi feito baseado em seqüências do projeto Genoma Café. Após detectar polimorfismos em uma pequena amostra de sete genótipos pertencentes a grupos genéticos diferentes (Congolese SG1, Congolese SG2, Congolese B, Congolese C, Guineans, Uganda selvagem e Uganda N'Ganda), uma amostra maior de genótipos pertencentes estes grupos genéticos foi analisada para os dois genes que parecem ter os maiores papeis na biossintese da sacarose no gênero Coffea (SUS1, SUS2). Genótipos de *C c,* pertencentes aos grupos genéticos de Congo e da Guiné foram obtidos do CNRA (Centre National de Recherche Agronomique) da Costa do Marfim, enquanto os genótipos do grupo Uganda (Uganda selvagem e N'Ganda) foram obtidos do NARO-CORI, Uganda. Todos os genótipos usados para análise interespecífica (14 espécies) foram obtidos junto à coleção ao IRD, em Montpellier, França.

Resultados e Discussão

Análise *in silico*

A mineração de seqüências nos bancos de dados pesquisados permitiu a identificação de genes para todas as enzimas estudadas. Para a sacarose sintase (SUS), a sacarose fosfato sintase (SPS) e invertase de vacúolo (VI) foram encontrados dois genes para cada enzima. Para a enzima invertase de parede (CWI) foram encontrados sete genes.

Para a SUS1, a primeira isoforma da sacarose sintase, foram localizadas 259 seqüências com 69 polimorfismos entre elas, sendo 62 do tipo SNP e sete INDELs ou SSR. Quarenta e três destes ficam na região codante, sendo que apenas sete induzem uma modificação da proteína (16,2%). Na segunda isoforma da sacarose sintase (SUS2) foram encontrados apenas 19 polimorfismos em 23 seqüências, sendo todos os polimorfismos do tipo SNP. Todos foram encontrados na região que codifica para a proteína e cinco eram do tipo não sinônimo (NS), isto é alteram a seqüência protéica (25 %).

Para as invertases de parede foram encontrados sete *contigs*, mas só para dois deles o número de seqüências (4 e 6 seqüências respectivamente) disponíveis permitiu a busca de polimorfismos, sendo que somente um polimorfismo foi identificado. Para a invertase vacuolar, só um *contig* foi analisado (para o outro só um singleton está disponível em todos os bancos de ESTs analisados). Para esse *contig* foi possível detectar 20 polimorfismos do tipo SNP nas oito seqüências

encontradas. Destes polimorfismos encontrados, oito (40%) eram não sinônimos (NS).

Sobre a via de biossíntese dos diterpenos, foram analisadas enzimas envolvidas na via geral dos isoprenoides (IDS, DXR, MPDC, MECPS, HMGR) e também enzimas da via de biossintese dos caurenos (CPS, KS, KO). Para a enzima IDS, que é envolvida na via plastidial MEP, foram encontradas 53 seqüências, sendo esta a enzima que apresentou uma das maiores taxas de alteração da estrutura primária da proteína . Das seqüências encontradas nos bancos de dados estudados, foram detectados 27 polimorfismos: 24 do tipo SNP, dois do tipo indel e um do tipo SSR. Na região codante do gene foram encontrados 23 polimorfismos, sendo 56% não sinônimos. Para a enzima DXR, os bancos de dados possuem 24 seqüências que, alinhadas, permitiram detectar 33 polimorfismos, onde 29 eram do tipo SNP. Dos 30 polimorfismos presentes na região codante do gene, apenas sete (23%) eram não-sinônimos. As enzimas MPDC, MEPCS e HMGR apresentaram *contigs* de pequeno tamanho e, por conseqüência, pequeno número de polimorfismos, sendo na sua maioria sinônimos (S).

Entre as diferentes enzimas estudadas da rota dos caurenos, a enzima KO apresentou duas isoformas. A isosforma um (KO1) apresentou 82 seqüências de bases com 41 polimorfismos, sendo 40 do tipo SNP. Destas, 24 estavam presente em regiões codantes do gene e oito (33%), causavam alteração na proteína. A outra isoforma (KO2) apresentou dez seqüências com oito polimorfismos, todos do tipo SNP. Destes oito polimorfismos, quatro estavam na região que codificava a proteína e destes quatro, 50% era não sinônimo. Dos polimorfismos apresentados pela enzima CPS, 39 eram do tipo SNP, três do tipo indel e um do tipo SSR. Destes, 40 estavam na região que codificavam a proteína e a metade dos polimorfismos eram não-sinônimos. Para a enzima KS foram encontradas 31 seqüências nos bancos de dados, o que possibilitou a detecção de 38 polimorfismos, sendo todos do tipo SNP. Destes, 25 estavam presentes na região que codifica o gene e 60% eram não-sinônimos.

Analise *in vivo* por seqüênciamento de genes envolvidos na biossíntese de sacarose

Para a via da biossíntese da sacarose, foram seqüenciados genes de pelo menos um genótipo de cada grupo de diversidade de *Coffea canephora* e um genótipo de 14 outras espécies de *Coffea*. Os genes da SUS1 e SUS2 foram analisados utilizando um numero maior de genótipos.

Em Cc, um total de 13.000 pb foi analisado, detectando 181 polimorfismos, sendo 157 SNP, 11 indels e 13 SSR, com uma média de 1,2 SNP para cada 100 pb. Quando considerados todos os fragmentos de seqüências, a maioria dos SNP encontrava-se em regiões não traduzida dos genes (69%). Um terço dos polimorfismos detectados na região codante do gene foram não sinônimo (33 sinônimos e 16 não sinônimos).

As análises inter-específicas para a enzima SUS1, detectaram 100 polimorfismos, sendo 56 na região de exons. Dos 56 polimorfismos encontrados, 19 foram em espécies que não estão presentes nestes bancos e se estão, não possuem seqüências das mesmas regiões onde foram detectados os polimorfismos *in silico*. Os 37 polimorfismos restantes estavam localizados em seqüências de espécies presentes nos bancos de dados (*C. arabica*, *C. canephora* e *C. racemosa*) e todos foram localizados confirmando o polimorfismo detectado *in silico*. Na análise *in vivo* das seqüências de *Cc*, foram detectados 36 polimorfismos, sendo que 28 estavam presentes na região codante do gene. Destes, apenas nove (26%) eram polimorfismos presentes nos bancos de dados.

Entre as diferentes espécies estudadas, foi possível detectar 139 polimorfismos para SUS2. Apenas 33 em região de exons e, destes, 25 eram polimorfismos pertencentes a outras espécies que não aquelas presentes nos bancos de dados ou, se pertenciam a essas espécies, as seqüências dessa região não estão disponíveis. Dessa forma, foram encontrados apenas sete polimorfismos que estão presentes nos três bancos de dados. Quando foi analisada apenas a espécie *C. canephora*, foram encontrados 97 polimorfismos no total. Destes, 42 estavam presentes em regiões de exons, entretanto seis polimorfismos estavam presentes nas seqüências *in vivo* e *in silico.*

A busca de polimorfismos dentro dos bancos de ESTs disponíveis permitiu a identificação de 317 polimorfismos dentro dos 14 genes estudados. Uma media de 1.1 polimorfismos foram encontrados para cada 100 bp.

Uma alta variabilidade do número de polimorfismos foi observada entre os diferentes genes, de zero para a CWI1 a 69 para SUS1. Essa variabilidade depende do número de ESTs presentes nos banco de dados, das regiões disponíveis (maior número de polimorfismos nas regiões não traduzidas) e também da velocidade de evolução dos genes.

A análise dos tipos de polimorfismos detectados nas regiões que codificam para as proteínas revelou uma variabilidade da relação de polimorfismos S e NS entre os genes. Os genes envolvidos na biossíntese da sacarose apresentaram uma relação menor (em média) de polimorfismo significativo e não significativo que os genes dos diterpenos, significando que esses genes parecem estar sob alta limitação evolutiva. Isso faz sentido em razão do papel central do metabolismo da sacarose para o crescimento das plantas e sua implicação na resposta aos estresses.

A comparação dos resultados obtidos *in silico* e *in vivo* para SUS1 e SUS2 permitiu verificar a eficiência da busca *in silico* dos polimorfismos. Esses resultados obtidos *in silico* podem também ser usados para estudar a evolução dos genes e definir suas importâncias evolutivas. Além disso, ao contrario da busca *in vivo* que está restrita a alguns genes de interesse e com custo muito elevado, a estratégia *in silico* pode oferecer rapidamente uma imagem do nível de polimorfismo no genoma inteiro, porém as duas estratégias são complementares.

Uma análise mais aprofundada será feita sobre os polimorfismos encontrados em *C. arabica* para definir suas origens, pois sendo essa espécie alotetraploide haverá polimorfismos oriundos dos dois genomas parentais (*C. canephora* e *C. eugenioides*). A comparação dos dados obtidos a partir das seqüências de Cc e Ca deve permitir a identificação dos tipos de polimorfismos: os correspondendo a polimorfismos fixados entre as espécies parentais e os polimorfismos que segregam dentro de cada genoma.

Agradecimentos

Ao Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café, ao CNPq, a Embaixada da França no Brasil pelo apoio financeiro dos pesquisadores D. Pot e P. Marraccini e a equipe do "Genomic and Coffee Quality" do IRD (Montpellier, França), especialmente ao Dr Alexandre de Kochko, pela cessão de material vegetal para realização deste estudo.

Referências bibliográficas

Lin, C.; Mueller, L.A.; Mc Carthy, J.; Crouzillat, D.; Pétiard, V.; Tanksley, S.D. (2005) Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts. *Theor Appl Genet*. 112:114-130.

Vieira, L.G.E.; Andrade, A.C.; Colombo, C.A.; Moraes, A.H.de.A.; Metha, A.; Carvalho de Oliveira, A.; Labate, C.A.; Marino, C.L.; Monteiro-Vitorello, C.de.B.; Monte, D. de C.; Giglioti, E.; Kimura, E.T.; Romano, E.; Kuramae, E.E.; Lemos, E.G.M.; Pereira de Almeida, E.R.; Jorge, E.C.; Albuquerque, E.V.S.; da Silva, F.R.; Vinecky, F.; Sawazaki, H.E.; Dorry, H.F.A.; Carrer, H.; Abreu, I.N.; Batista, J.A.N.; Teixeira, J.B.; Kitajima, J.P.; Xavier, K.G.; Maria de Lima, L.; Aranha de Camargo, L.E.; Pereira, L.F.P.; Coutinho, L.L.; Lemos, M.V.F.; Romano, M.R.; Machado, M.A.; Costa, M.M. do. C.; Grossi, de Sá M.F.; Goldman, M.H.S.; Ferro, M.I.T.; Tinoco, M.L.P.; Oliveira, M.C.; Van Sluys, M-A.; Shimizu, M.M.; Maluf, M.P.; Souza da Eira, M.T.; Guerreiro Filho, O.; Arruda, P.; Mazzafera, P.; Mariani, P.D.S.C.; de Oliveira, R.L.B.C.; Harakava, R.; Balbao, S.F.; Tsai, S.M.; di Mauro, S.M.Z.; Santos, S.N.; Siqueira, W.J.; Costa, G.G.L.; Formighieri, E.F.; Carazzolle, M.F.; Pereira, G.A.G. (2006). Brazilian coffee genome project: an EST-based genomic resource. *Braz. J. Plant Physiol*., 18(1):95-108.