IDENTIFICAÇÃO DE POLIMORFISMOS DE NUCLEOTÍDEOS ÚNICOS EM MONTAGEM DE ESTs DE TRÊS ESPÉCIES DE CAFÉ

Ramon O. VIDAL[1], E-mail: vidal@lge.ibi.unicamp.br; Marcelo F. CARAZZOLLE[1]; Cláudio L. M. SAMPAIO[1]; Gustavo G. L. COSTA[1]; Eduardo F. FORMIGHIERI[1]; David POT[2]; Jorge M. C. MONDEGO[1]; Gonçalo A. G. PEREIRA[1]

[1]Laboratório de Genômica e Expressão, Instituto de Biologia, UNICAMP, 13083-970, Campinas, SP, Brazil; [2]Cirad, UMR PIA, Avenue d'Agropolis, F-34398 Montpellier Cedex 5, France

Resumo:

A partir do seqüenciamento de ESTS de três espécies de café, *Coffea arabica*, *Coffea canephora* e *Coffea racemosa*, foi aberta uma série de possibilidades para o estudo de características diferenciais entre essas espécies. Características fenotípicas de interesse agronômico podem ser estudadas molecularmente através da análise de transcritos dessas três espécies e também através do mapeamento de um conjunto de genes alvos para programas de melhoramento genético. Neste trabalho foi realizado um agrupamento de ESTs das três espécies por similaridade visando estudo de polimorfismos de nucleotídeos únicos (SNPs). A partir de parâmetros de similaridade (>95%) e sobreposição (>100 bp) foram construídos 15.885 *contigs*, e nestes observa-se uma freqüência de 0,47 SNPs a cada 100 pb. Os resultados foram armazenados para consulta visando futuros estudos de variabilidade em indivíduos de cada espécie e entre estas três espécies.

Palavras-chave: *Coffea, EST, SNP, transcriptome, CAP3*

IDENTIFICATION OF SINGLE NUCLEOTIDE POLYMORPHISMS IN ASSEMBLY OF THREE COFFEA SPECIES ESTs

Abstract:

The sequencing of the ESTS of three coffee species transcriptome, Coffea arabica, Coffea canephora and Coffea racemosa, opened many possibilities for studying different characteristics between these species. Fenotipical characteristics of agronomic interest can be molecularly studied through the transcript analysis of these three species and also through the mapping of a set of target genes for genetic improvement programs. In this work, a clustering for similarity of ESTs of the three species was carried, aiming a study of single nucleotide polymorphisms (SNPs). Using parameters of similarity (>95%) and overlapping (>100 bp) 15,885 contigs have been constructed, which contain a frequency of 0,47 SNPs for 100 bp. The results were stored for consults and future studies of variability in individuals of each species and between these three species.

Key words: Coffea, EST, SNP, transcriptome, CAP3

Introdução

O café é considerado um dos mais importantes produtos agrícolas no mercado internacional. . Existem, cerca de cem espécies do gênero *Coffea* descritas (Fazuoli, 1986), mas apenas duas espécies têm grande importância para o cultivo: *Coffea arabica* e *Coffea canephora*. Aproximadamente 70% do café comercializado mundialmente é do tipo arábica e o restante é de café robusta (*C. canephora*).

Apesar dos esforços contínuos, o progresso no implemento na produção do café através de abordagens convencionais tem sido muito lento, devido a diversos fatores tais como a estreita base genética do café cultivado e a falta de marcadores genéticos. Busca-se melhorias em características agronômicas tais como: florescimento, rendimento, tamanho do grão, qualidade da bebida, índice de cafeína, resistência a doenças e pragas e tolerância ao estresse hídrico.

O desenvolvimento de novas tecnologias aplicadas à biologia vem gerando um vasto conhecimento na área de genômica de plantas. O seqüenciamento em larga escala de cDNA visando a produção de ESTs (*Expressed sequenced tags*) fornece evidências diretas para todas as amostra de transcritos de um genoma, permitindo a rápida caracterização do conjunto do genes expressos.

A espécie *C. canephora* foi seqüenciada pela iniciativa *Nestlé Research Center* (Lin, 2005), enquanto as espécies *C. arabica e C. racemosa* foram seqüenciadas pelo *Brazilian Coffee Genome Project*, formulado em 2002 através de um acordo cooperativo assinado entre o *Brazilian Coffee Research* e o *Development Consortion* (CBP&D-Café), um consórcio nacional de 40 universidades públicas e institutos de pesquisa. O projeto conta também com a participação da Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), da Fundação de Amparo à Pesquisa na Bahia (FAPESP) e do Fórum Permanente das Relações Universidade-Empresa (UNIEMP). O Laboratório de Genômica e Expressão (LGE -

http://www.lge.ibi.unicamp.br) da Universidade Estadual de Campinas (UNICAMP) foi designado como o centro da bioinformática para armazenamento dos bancos de dados de seqüências do Genoma do Café e como coordenador de todos os aspectos relacionados com a submissão de seqüências, performance e produtividade de todos os grupos de seqüenciamento, armazenamento de informações, análise comparativa através do programa de alinhamento local BLAST e clusterização dos ESTs.

Uma das formas de estudar diferenças entre os indivíduos de uma mesma espécie e entre as espécies é através da detecção e análise dos polimorfismos de nucleotídeos únicos (SNPs) e de inserções e deleções (INDELs). Um SNP é uma única mudança de nucleotídeo numa mesma posição de molécula de DNA entre indivíduos, diferindo de mudanças em múltiplas bases em posições aleatórias. SNPs podem ser responsáveis por importantes variações nas características fenotípicas entre indivíduos de uma mesma espécie (Emahazion *et al, 2001*; Sherry *et al,* 2001).Os métodos de detecção de SNPs *in silico* buscam por diferenças individuais em uma seqüência através de analise computacional (Buetow et al, 1999; Marth et al, 1999; Picoult-Newberg et al, 1999), tendo a capacidade de discernir entre um verdadeiro polimorfismo e erros de seqüenciamento.

O objeto de estudo deste trabalho é mapear possíveis SNPs que ocorrem em genes do café a partir de uma montagem de ESTs das três espécies supracitadas..

Material e Métodos

Foram utilizadas 275.117 seqüências de ESTs das três espécies de café, originadas a partir de 56 bibliotecas de diversos tecidos (Figura 1) e órgãos em várias condições. A espécie *Coffea arabica* representa 65% desse total, enquanto *C. canephora e C. racemosa* representaram 31% e 2%, respectivamente. Essas seqüências foram tratadas com o programa LUCY para retirar a seqüências de vetor, caudas PolyA/T e seqüências de baixa qualidade.

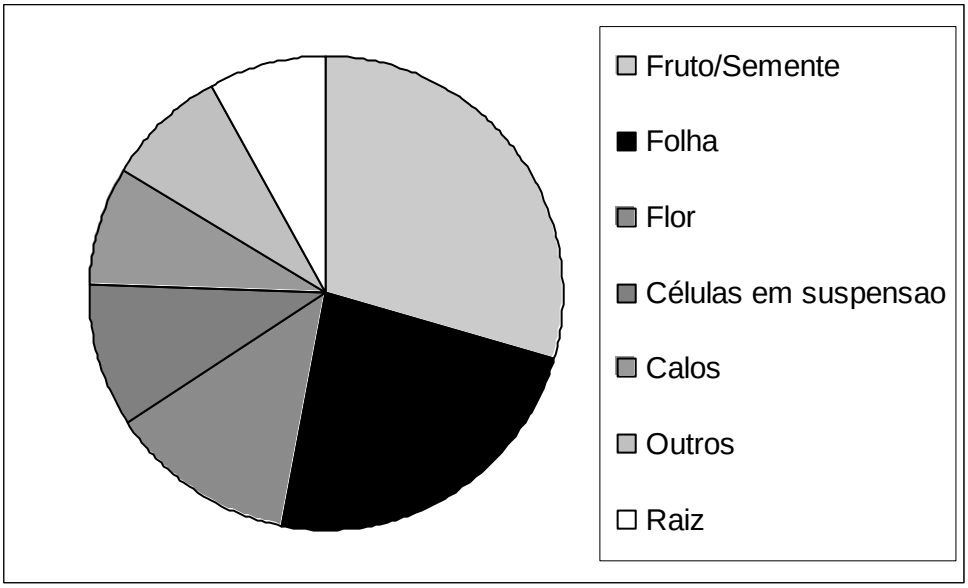

Figura 1: Distribuição de biblioteca por tecido

A montagem dos ESTs de café foram realizadas através do programa CAP3. Os parâmetros de alinhamento foram ajustados buscando reduzir o número de *contigs* e *singlets*, evitando que genes diferentes que compartilham de regiões em comum se agrupem num mesmo *contig*. O arquivo ACE gerado pelo CAP3 foi usado como entrada do programa AutoSnp que encontra todas as discrepâncias com alta qualidade (reduzindo os falsos positivos por erros de seqüenciamento) nas montagens, o que geralmente está associado com a presença de polimorfismos. Todo o processo foi automatizado para o estudo em larga escala. Os *scripts* para manipulação de dados e geração de relatórios foram escritos em PERL.

Resultados e Discussão

Depois do processo de trimagem, os 194.455 ESTs foram utilizados como entrada para o processo de montagem global. ESTs individuais foram montados com o programa CAP3 utilizando seis diferentes conjuntos de parâmetros (Tabela 2), que demonstraram resultados muito próximos no número de *contigs* e *singlets*. A melhor montagem para o nosso propósito foi a que alinhou regiões de ESTs compartilhando ao menos 95% de identidade de seqüência com >100 bp de sobreposição entre eles (>20% do tamanho médio dos EST após a etapa de limpeza com o LUCY). Como resultado foram obtidos 15.885 *contigs* e 47.361 *singlets* em 276,6 Mb de seqüências transcritas com um tamanho médio de 599 pb (DP = 309). Uma análise de similaridade entre as seqüências dessa montagem resultou na identificação de mais de 30 mil "genes únicos" (unigene), entre *contigs* e *singlets*. Esta informação básica fornece um valioso recurso a ser utilizado e minerado para estudos na fisiologia das plantas do café, o que ajudará no isolamento e caracterização de genes com importância agronômica.

Tabela 1: ESTs analisados totais e trimados por alta qualidade

Espécie	Total de ESTs	%	ESTs trimados com alta qualidade *	%
CA	189.351	68,83	128.139	65,90
CC	78.182	28,42	61.001	31,37
CR	7.584	2,76	5.315	2,73
Total	275.117		194.455	

*Seqüências foram tratadas com o programa LUCY (ftp://ftp.tigr.org/pub/software/Lucy)

Aproximadamente 83% dos *contigs* (13.220) tiveram seus ESTs com origem de mais de uma biblioteca, e 62% por mais de uma espécie (Tabela 2). A figura 2 mostra que a maioria dos *contigs* formados por mais de três ESTs são misturas de mais de uma espécie e que esse valor de *contigs* mistos aumenta conforme a quantidade de ESTs que os formam. Esses resultados sugerem que variações nucleotídicas entre as espécies têm somente um pequeno efeito na montagem global e que a presença de SNPs em contigs formados por múltiplos ESTs provavelmente estão presentes entre espécies.

Nestes 15.885 *contigs* foram detectados 55.171 SNPs (30.572 transições e 25.499 transversões) e 12.438 indels) no total. A partir desses resultados, a freqüência calculada da ocorrência dos SNPs ao longo das seqüências é de 0,47/100 pb. Os resultados obtidos estão armazenadas no servidor do LGE para consulta dos usuários cadastrados e para uma futura análise que será feita visando a classificação dos SNPs como sinônimos (altera o aminoácido a ser codificado) ou não-sinônimos (não altera o aminoácido a ser codificado) e a associação destes SNPs às famílias gênicas mais abundantes.

Tabela 2: Resumo da montagem dos *contigs* de café por parâmetro utilizado

Parâmetros do CAP3	Quantidade de *contigs*	Tamanho médio dos *contigs*	*Contigs* contendo ESTs de ao menos duas espécies	*Contigs* contendo ESTs de varias bibliotecas	*Contigs* contendo ESTs das três espécies	*Contigs* com ao menos uma seqüência de cada espécie		
						CA	CC	CR
-o 50[a] – p 90[b]	15,584	955	65,14%	85,29%	10,90%	86,00%	74,01%	16,04%
-o 50 – p 95	15,584	955	65,14%	85,29%	10,90%	86,00%	74,01%	16,04%
-o 100 – p 90	**15,603**	**937**	**64,35%**	**84,73%**	**10,52%**	**85,60%**	**73,67%**	**15,61%**
-o 100 – p 95	**15,885**	**917**	**62,24%**	**83,34%**	**9,35%**	**84,49%**	**72,44%**	**14,66%**
-o 150 – p 90	15,500	919	63,64%	84,23%	10,05%	85,29%	73,15%	15,24%
-o 150 – p 95	15,738	900	61,53%	82,90%	8,95%	84,15%	72,07%	14,26%

a – tamanho da região sobreposta
b – porcentagem de similaridade

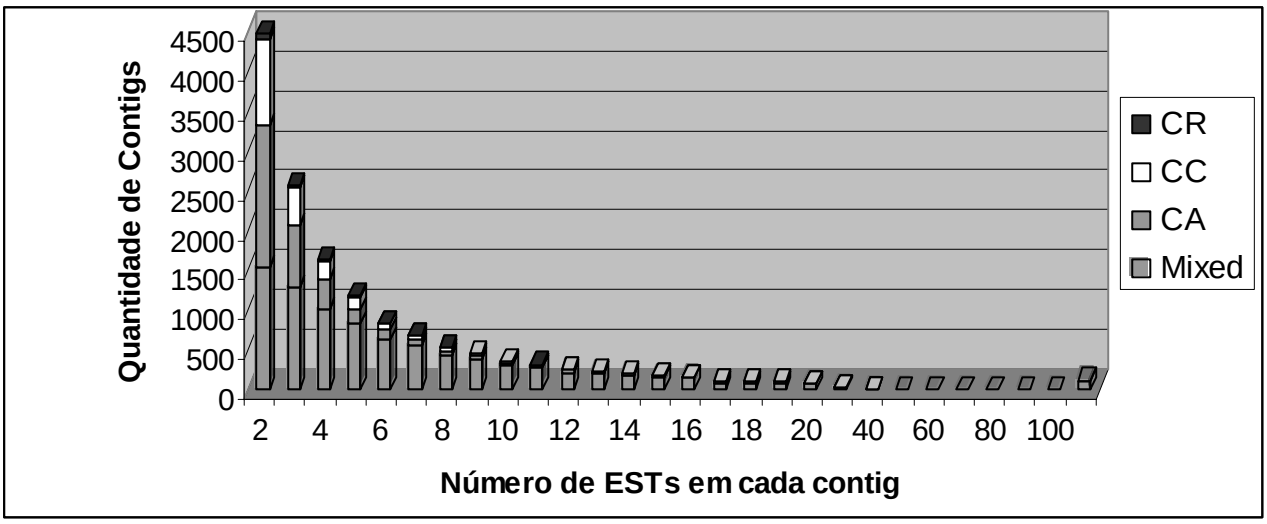

Figura 2. Número de ESTs (eixo X) por quantidade de *contigs* (eixo Y). A legenda descreve o tipo de contig formado pela espécie *C. Racemosa (CR), C. Canephora (CC), C. Arabica (CA)* e *contigs* formados por ESTs originados por mais de uma destas espécies (*Mixed*).

Conclusões

O pré-processamento das seqüências é um passo importante para evitar que seqüências de má qualidade atrapalhem o correto agrupamento dos ESTs em contigs. Esse agrupamento dos transcriptomas das três espécies e a mineração de informação proveniente dos contigs e singlets será um valioso recurso para a compreensão de características estruturais e funcionais nos genes expressos entre os indivíduos, diferentes espécies e diferentes bibliotecas do café.

A quantidade de genes únicos das três espécies é bastante elevada levando em conta outras espécies de planta como arroz, arabidopsis e milho que estão por volta de 30 mil, porém como estamos lidando com uma montagem de três espécies diferentes é muito provável que muitos genes não possuam similaridade suficientemente grande para se alinharem num gene único e acabam se tornando singlets. Contando apenas os genes da espécie com maior numero de ESTs (*Coffea arabica*) podemos chegar no máximo a 40.000 genes, o que pode ser reduzido levando em conta que ainda podem existir erros comuns de montagem.

A quantidade de SNPs foi superestimada devido aos contigs formados por apenas dois ESTs, as discrepâncias nestes contigs não são estatisticamente confiáveis, porém SNPs em contigs maiores, com discrepâncias numa mesma posição, é um maior indicativo de que seja real. Porém o valor estimado para o café de 0.47 SNPs a cada 100 bp nestas condições é próxima da freqüência do arroz (0.43 por 100 bp) (Feltus *et al*, 2004). O conjunto de dados de SNPs (disponível em http://www.lge.ibi.unicamp.br/cafe) é um valioso recurso para experimentos genéticos envolvendo o café e suas diversas espécies e indivíduos.

Referências Bibliográficas

Buetow, K.H., Edmonson, M.N., and Cassidy, A.B., Reliable identification of large numbers of candidate SNPs from public EST data, Nat. Genet., 21:323–325, 1999.

Emahazion T., Feuk L., Jobs M., Sawyer SL., Fredman D., St Clair D,; Prince JA., Brookes AJ. SNP association studies in Alzheimer's disease highlight problems for complex disease analysis. Trends in Genética. 17:407-413, 2001

Fazuoli, L.C. Genética e melhoramento do cafeeiro. In: Rena,A.B.; Malavolta,E.; Rocha,M.; Yamada,T. (eds). *Cultura do cafeeiro* – fatores que afetam a produtividade. Piracicaba, Associação Brasileira para Pesquisa da potassa e do fosfato. 87-113. 1986

Feltus F.A. et al. An SNP resource for rice genetics and breeding based on subspecies Indica and Japonica genome alignments. Genome research 14(9): 1812-1819. 2004

Lin C, Mueller LA, Carthy JMc. Coffee and tomato common gene repertoires as revealed by deep sequencing of seed and cherry transcripts. Theor Appl Genet 112: 114-130. 2005

Marth, G.T. et al. A general approach to single-nucleotide polymorphism discovery, Nat. Genet., 23(4):452–456, 1999.

Picoult-Newberg, L. et al., Mining SNPs from EST databases, Genome Res., 9(2):167–174, 1999.

Sherry ST, Ward MH , Kholodov M, Baker J, Phan L, Smigielski EM, and Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucl. Acids Res. 29: 308-311. 2001