

OBTENÇÃO DE ESTs POTENCIALMENTE RELACIONADAS À RESISTÊNCIA DO CAFEIEIRO À FERRUGEM POR MEIO DE ANÁLISE *IN SILICO*

Flávia THIEBAUT Andrade¹, E-mail: biocafe@ufv.br; Eveline Teixeira CAIXETA^{1,2}; Bárbara HUFNAGEL Maciel¹; Samuel Mazzinghy ALVARENGA¹; Eunize Maciel ZAMBOLIM^{1,3}; Ney Sussumu SAKIYAMA^{1,4}

¹ Universidade Federal de Viçosa (UFV)/ BIOAGRO, Laboratório de Biotecnologia do Cafeeiro (Biocafê), 36570-000, Viçosa-MG. ²Embrapa Café,

³ UFV/ Departamento de Fitopatologia, ⁴ UFV/ Departamento de Fitotecnia. Apoio Financeiro: PNP&D/CAFÊ

Resumo:

A partir das informações geradas no Projeto Brasileiro do Genoma Café (PBGC) foram identificadas, por meio de análise *in silico*, seqüências potencialmente envolvidas na resistência a *Hemileia vastatrix* Berk. et Br. presentes nos cafeeiros. Foram usadas diferentes estratégias para minerar as seqüências de interesse. Inicialmente, palavras-chave que correspondem a termos relacionados aos mecanismos de resistência de plantas a patógenos foram obtidas da literatura e utilizadas como “iscas” para mineração dos dados. Com o auxílio de ferramentas disponíveis na plataforma de bioinformática do PBGC, criaram-se projetos englobando as ESTs (*Expressed Sequence Tags*) relacionadas a cada uma destas palavras. Outra estratégia utilizada foi a busca por similaridades entre algumas seqüências envolvidas com a resistência do cafeeiro a doenças já publicadas com as seqüências do PBGC, por meio do algoritmo BLAST. Utilizou-se, também, o “*Electronic Northern*”, uma ferramenta desenvolvida pelo Laboratório de Genômica e Expressão (LGE). A mineração, usando as três estratégias, identificou 8.968 seqüências do PBGC. O envolvimento ou ligação destas seqüências com genes de resistência à ferrugem do cafeeiro será confirmado, em trabalhos futuros, utilizando marcadores moleculares e técnicas de genômica funcional.

Palavras-chave: mineração de dados, café, bioinformática, genômica.

OBTAINING ESTS POTENTIALLY RELATED WITH COFFEE TREE RESISTANCE TO LEAF RUST BY *IN SILICO* ANALYSIS

Abstract:

In silico analyses using information generated by the Projeto Brasileiro do Genoma Café (PBGC) permitted the identification of sequences potentially involved with coffee tree resistance to leaf rust. These analyses aimed the generation of knowledge concerning the coffee tree resistance factors to *Hemileia vastatrix* Berk. et Br. Three different strategies for the identification of the sequences related to the resistance were used. Initially, key-words correspondent to plants resistance mechanism to pathogens were obtained and used as "baits" for data mining. Projects including the ESTs (*Expressed Sequence Tags*) related to each one of these words were created with the help of the available tools at the PBGC bioinformatics platform. The search for similarities between some published sequences and sequences from the PBGC, by using the BLAST algorithm was another used strategy. It was also used the *Electronic Northern*, a tool developed by the Laboratório de Genômica e Expressão (LGE). Eighty nine hundreds and sixty eight sequences of the PBGC were identified by data mining. The relationship or linkage of these sequences with coffee leaf rust resistance genes will be further confirmed by molecular markers and functional genomics techniques.

Key words: data mining, coffee, bioinformatics, genomics.

Introdução

Os grandes avanços da biologia molecular e da bioinformática deram origem a era da genômica que visa estudar o genoma como um todo. Atualmente, já estão disponíveis as seqüências do genoma de 520 organismos, incluindo bactérias, leveduras, fungos e algumas plantas superiores e estão em andamento 1.914 projetos genomas (<http://www.genomesonline.org/gold.cgi>). O seqüenciamento do genoma de várias plantas tem facilitado e acelerado a identificação de genes responsáveis por características agrônomicas desejáveis, possibilitando a manipulação subsequente de genes de interesse por meio de técnicas de genética molecular (Souza e Silva, 2002).

A cultura do café também está sendo beneficiada pela revolução genômica. No Projeto Brasileiro do Genoma Café (PBGC) foram seqüenciados cerca de 200.000 ESTs (*Expressed Sequence Tags*), provenientes das espécies *Coffea arabica*, *C. canephora* e *C. racemosa*. Estas ESTs foram agrupadas, resultando em 17.982 *contigs* e 32.155 *singlets*. Todos estes dados, além de várias ferramentas genéticas e genômicas que auxiliam os estudos e pesquisas, foram disponibilizados pelo Laboratório de Genômica e Expressão (LGE - <http://www.lge.ibi.unicamp.br/cafe>) (Vieira et. al., 2006).

Desta forma, este trabalho tem como objetivo identificar, por meio da análise *in silico*, seqüências do banco de dados do PBGC envolvidas com a resistência do cafeeiro à ferrugem, que é a principal doença desta cultura. Estas ESTs poderão ser úteis para desenvolver novos e eficientes marcadores moleculares, que serão usados para assistirem os programas de melhoramento, bem como para gerar informações que visam direcionar as manipulações genéticas nos programas. **Material e Métodos**

Para a mineração das seqüências potencialmente envolvidas com a resistência do cafeeiro a patógenos, analisou-se as ESTs do Projeto Brasileiro do Genoma Café (PBGC) por meio da plataforma de bioinformática do LGE (<http://www.lge.ibi.unicamp.br/cafe/>). Foram utilizadas três estratégias de mineração.

Na primeira estratégia usou-se como “iscas”, palavras-chave que estão relacionadas com resistência, de acordo com a literatura. As palavras-chave utilizadas foram: LRR, NBS, *pathogenesis* (PR), *thaumatin*, *chitinase*, HSP, *hypersensitive*, *polyphenoloxidase*, *importin*, *glucanase*, *phytoalexin*, *chalcone*, *resistance*. Com estas “iscas” e por meio do aplicativo *Gene Project*, foram construídos 13 projetos buscando dentro do banco de dados do Genoma os *reads* de interesse. Os *reads* de cada projeto foram agrupados e os *contigs* formados foram submetidos a uma busca por similaridade com bancos de dados, através do algoritmo BLAST.

Na segunda estratégia foi realizada uma busca por similaridade de seqüências publicadas em artigos científicos com as seqüências do PBGC. Conforme as literaturas pesquisadas (Chen et al., 2003; Fernandez et al., 2004; Lin et al., 2005; Ganesh et al., 2006;) as seqüências codificam genes de cafeeiros relacionados à resistência. Estas foram submetidas a uma busca por similaridade com as ESTs depositadas na base de dados do PBGC.

A outra estratégia foi baseada nos dados fornecidos pelo aplicativo “*Electronic Northern*”. Esta ferramenta, desenvolvida pelo LGE, agrupa todas as 200.000 ESTs descrevendo todos os *contigs* e *singlets* formados e a porcentagem de *reads* existentes em cada uma das 37 bibliotecas. Foram selecionados *clusters* que continham *reads* provenientes da biblioteca RM1 (biblioteca de folhas de *Coffea arabica* infectadas com bicho mineiro e ferrugem). Nesta biblioteca podem-se encontrar genes relacionados com a resistência à ferrugem.

Resultados e Discussão

Para mineração das ESTs, inicialmente, foi realizada uma busca na literatura por palavras-chave, bem como pesquisado a sua relação no processo de resistência. As seguintes palavras-chave e seus respectivos projetos formados foram:

- LRR e NBS, que são domínios conhecidos como relacionados a genes de resistência a patógenos, formaram um projeto LRR contendo 825 *reads*, que após o agrupamento resultou em 160 *contigs* e 243 *singlets*, e o projeto NBS com 494 *reads* sendo 105 *contigs* e 142 *singlets*.

- *Pathogenesis* refere-se a uma proteína relacionada com a patogênese. A partir desta palavra-chave, criou-se um projeto com 979 *reads*. Após o agrupamento foram obtidos 63 *contigs* e 37 *singlets*.

- Por meio da palavra-chave *Thaumatin* foram obtidos 317 *reads* dos quais se formaram 16 *contigs* e sete *singlets*.

- *Chitinase* são proteínas altamente expressas no café. Seu modo de ação consiste no enfraquecimento da parede celular do fungo, resultando na lise celular e posterior morte do patógeno. O projeto criado a partir desta “isca” apresentou 1.855 *reads* que deram origem a 47 *contigs* e 48 *singlets*.

- HSP (*Heat Shock Protein*) são proteínas essenciais na via de transdução de sinais de defesa da planta. Com essa palavra-chave foi obtido um projeto contendo 240 *reads* que resultaram em 31 *contigs* e 27 *singlets*.

- A palavra *Hypersensitive* foi utilizada porque está relacionada à hipersensibilidade da planta. O projeto criado possui 86 *reads*, que resultou em apenas oito *contigs* e quatro *singlets*.

- *Polyphenoloxidase* está relacionada ao mecanismo de defesa a patógenos e a insetos. O projeto construído apresentou 67 *reads*, resultando em quatro *contigs* e três *singlets*.

- Com a palavra *Importin* foi criado um projeto com 532 *reads* que após o agrupamento resultaram em 11 *contigs* e quatro *singlets*.

- *Glucanase* refere-se a uma proteína relacionada com a patogênese, sendo que a sua atividade antifúngica na planta provavelmente está na hidrólise do β 1,3-glucan presente na parede celular do fungo, resultando em enfraquecimento da parede e, conseqüentemente, em lise e morte celular. O projeto obtido a partir desta palavra-chave contém 642 *reads*, sendo 92 *contigs* e 68 *singlets*.

- *Phytoalexin* é uma substância sintetizada por plantas quando atacadas por patógenos. O projeto formado consiste em 12 *reads* sendo apenas três *contigs* e dois *singlets*.

- *Chalcone* está envolvida na síntese de aminoácidos essenciais à planta. Construiu-se um projeto contendo 153 *reads* que formaram cinco *contigs* e oito *singlets*.

- Utilizando a palavra *Resistance* como “isca” foi possível construir um projeto contendo 1.864 *reads*. Após o agrupamento das seqüências mineradas foram obtidos 347 *contigs* e 416 *singlets*.

Estes resultados estão resumidos na figura 1 e tabela 1.

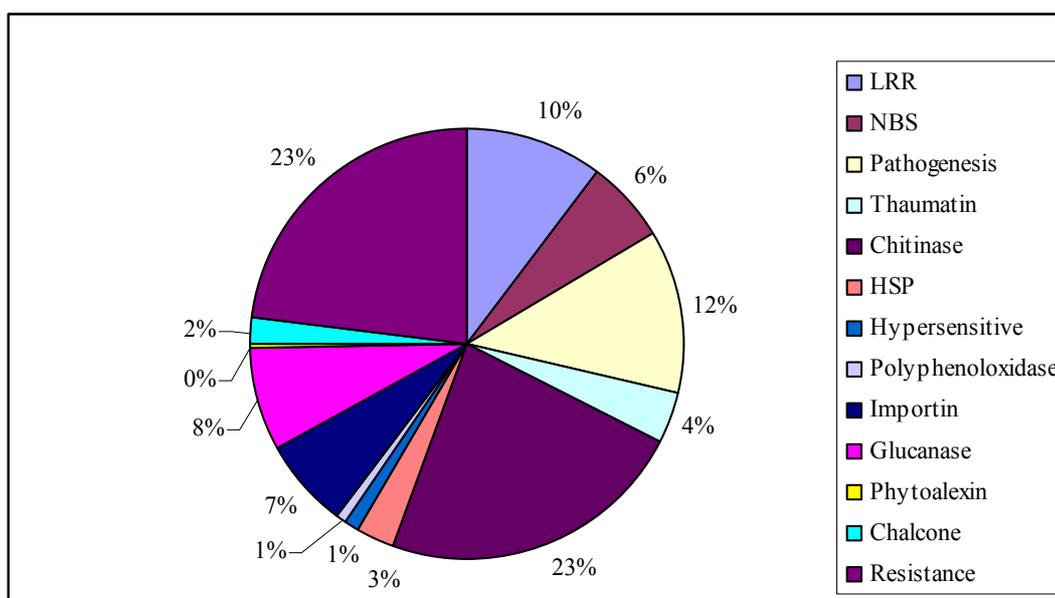


Figura 1: Percentagem dos reads obtidos pelos projetos obtidos com palavras-chave (primeira estratégia)

Tabela 1: Número de seqüências obtidas por meio do aplicativo *Gene Projects* (primeira estratégia).

Palavras-Chave	<i>Reads</i>	<i>Contigs</i>	<i>Singlets</i>
LRR	825	160	243
NBS	494	105	142
Pathogenesis	979	63	37
Thaumatin	317	16	7
Chitinase	1855	47	48
HSP	240	31	27
Hypersensitive	86	8	4
Polyphenoloxidase	67	4	3
Importin	532	11	14
Glucanase	642	92	68
Phytoalexin	12	3	2
Chalcone	153	5	8
Resistance	1864	347	416
Total	8066	892	1019

Na segunda estratégia, através do código de acesso fornecido nos artigos científicos consultados, foram identificadas 66 seqüências que, com a ferramenta BLAST, foram utilizadas para minerar as ESTs correspondentes dentro do PBGC (Tabela 2).

Na terceira estratégia, utilizando-se o aplicativo “*Electronic Northern*”, foram minerados 2.206 *clusters* que continham *reads* da biblioteca RM1. Deste total, foram considerados de interesse para o trabalho 602 formados apenas por *reads* da biblioteca RM1, sendo 28 *contigs* e 574 *singlets*, 21 *clusters* formados por seqüências com 99 a 75% e 213 com 74 a 50% dos *reads* provenientes da RM1 (Tabela 2).

A mineração realizada, utilizando as três estratégias, resultou em 8.968 seqüências do PBGC que equivalem a 8.066 ESTs obtidas a partir da primeira estratégia (*Gene Projects*), 66 mineradas pela segunda estratégia (Seqüências Publicadas) e 836 pela terceira (*Electronic Northern*) (Figura 2).

Tabela 2: Número de seqüências mineradas utilizando a segunda a terceira estratégia.

		Número de seqüências mineradas	Total
Estratégia 2 - Seqüências Publicadas	Lin et al. (2005)	8	66
	Fernandez et al. (2004)	49	
	Ganesh et al. (2006)	3	
	Chen et al. (2003)	6	
Estratégia 3 - <i>Electronic Northern</i>	Seqüências com 100% dos reads	602	836
	Seqüências com 75% a 99% dos reads	21	
	Seqüências com 50% a 74% dos reads	213	

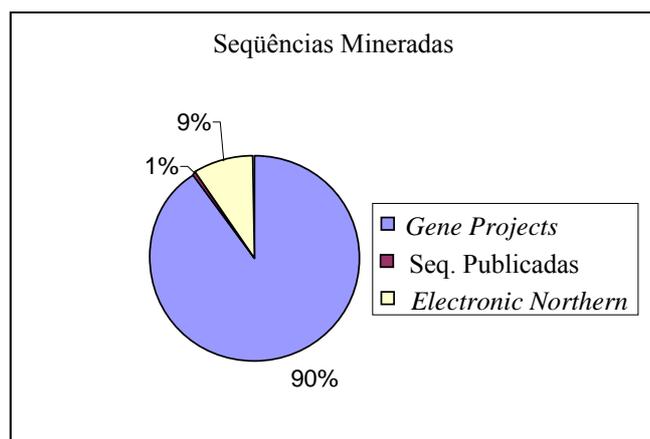


Figura 2: Valores totais de dados gerados na Mineração do Genoma do Café

Conclusões

Foram identificadas 8.968 seqüências do banco de dados do Projeto Brasileiro do Genoma Café potencialmente envolvidas com a resistência do cafeeiro a doenças. O envolvimento ou ligação destas seqüências com genes de resistência à ferrugem do cafeeiro será confirmado, em trabalhos futuros, utilizando marcadores moleculares e técnicas de genômica funcional.

Referências Bibliográficas

- Chen, Z. J.; Ribeiro, A.; Silva, M. C.; Santos, P.; Guerra-Guimarães, L.; Gouveia, M.; Fernandez, D.; Rodrigues Júnior, C. J. (2003) Heat shock-induced susceptibility of green coffee leaves and berries to *Colletotrichum gloeosporioides* and its association to *PR* and *hsp70* gene expression. *Physiological and Mol. Plant Pathol.*, 63(4):181-190.
- Fernandez, D.; Santos, P.; Agostini, C.; Bon, M.-C.; Petitot, A.-S.; Silva, M. C.; Guerra-Guimarães, L.; Ribeiro, A.; Argout, X; Nicole, M. (2004) Coffee (*Coffea arabica* L.) genes early expressed during infection by the rust fungus (*Hemileia vastatrix*). *Mol. Plant Pathol.*, 5(6):527-536.
- Ganesh, D.; Petitot, A.-S.; Silva, M. C.; Alary, R.; Lecouls, A.-C.; Fernandez, D. (2006) Monitoring of the early molecular resistance responses of coffee (*Coffea arabica* L.) to the rust fungus (*Hemileia vastatrix*) using real-time quantitative RT-PCR. *Plant Sci.*, 170(6):1045-1051.
- Lin, C.; Mueller, L. A.; Carthy, J. M.; 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(1):114-30.
- Souza, G. M.; Silva, A. M. (2002) SUCAST: Desvendando as vias de transdução de sinal da cana-de-açúcar. *Biotec. Ciência & Desenvolvimento*, 25:58-63.
- Vieira, L. G. E.; Andrade, A. C.; Colombo, C. A.; Moraes, A. H. de A.; Metha, A., et. al. (2006) Brazilian coffee genoma project: na EST-based genomic resource. *Braz. J. Plant Physiol.*, 18(1):95-108.