Short communication

Single nucleotide polymorphisms from *Theobroma cacao* expressed sequence tags associated with witches' broom disease in cacao

L.S. Lima[1], K.P. Gramacho[1], N. Carels[2], R. Novais[1], F.A. Gaiotto[2], U.V. Lopes[1], A.S. Gesteira[2], H.A. Zaidan[1,2], J.C.M. Cascardo[2], J.L. Pires[1] and F. Micheli[2,3]

[1]Centro de Pesquisas do Cacau, Itabuna, BA, Brasil
[2]Departamento de Ciências Biológicas, Universidade Estadual de Santa Cruz, Ilhéus, BA, Brasil
[3]Cirad-BIOS, UMR DAP, Montpellier, France

Corresponding author: F. Micheli
E-mail: fabienne.micheli@cirad.fr

ABSTRACT. In order to increase the efficiency of cacao tree resistance to witches' broom disease, which is caused by *Moniliophthora perniciosa* (Tricholomataceae), we looked for molecular markers that could help in the selection of resistant cacao genotypes. Among the different markers useful for developing marker-assisted selection, single nucleotide polymorphisms (SNPs) constitute the most common type of sequence difference between alleles and can be easily detected by *in silico* analysis from expressed sequence tag libraries. We report the first detection and analysis of SNPs from cacao-*M. perniciosa* interaction expressed sequence tags, using bioinformatics. Selection based on analysis of these SNPs should be useful for developing cacao varieties resistant to this devastating disease.

Key words: Single nucleotide polymorphisms; Expressed sequence tags; Bioinformatics; Witches' broom disease

Cacao (*Theobroma cacao* L.) is a cash crop grown by about two million producers, in more than 50 countries. The cacao tree is perennial and its growth may be characterized as intermittent. It has hermaphroditic flowers that are naturally pollinated by insects (primarily *Forcipomya* flies). Cacao is raised from seeds, which germinate and produce good plants when taken from pods not more than 15 days under ripe. Cacao multiplication is classically made through clone propagation (root-cutting). Cacao is challenged by many diseases, and witches' broom disease (WBD), caused by the fungus *Moniliophthora* (=*Crinipellis*) *perniciosa* (Stahel) Singer (Aime and Phillips-Mora, 2005), is one of the most important diseases of cacao. The disease is present in all the producing countries of Central and South America (Purdy and Schmidt, 1996). Host resistance is recommended as the best option for economic and sustainable control of WBD. The term "resistance" is used in cacao breeding programs in an epidemiological sense, which describes the ability to exclude or overcome, completely or to some degree, the effect of a pathogen (Agrios, 2005). For a perennial crop such as cacao, the substitution of susceptible varieties by resistant ones involves a considerable cost and requires a few years before economic production levels are obtained. In order to increase the durability of the resistance of cacao trees to WBD and to release this resistant material to the farmers, the use of clones carrying a great number of resistance genes constitutes the main objective of the cacao breeding program.

The construction of linkage maps allows the association of markers with agronomically important traits and subsequently marker-assisted selection (MAS), speeding up the breeding process (Tanksley et al., 1988). Selection can be even done in the absence of the expression of the phenotype. Another advantage of MAS is the accumulation of genes whose effect cannot be perceived based on the phenotype. For example, breeding for disease resistance is very desirable to accumulate (pyramiding) genes of resistance, in order to reduce the chances of pathogen breakdown of existing resistance. For this reason, obtaining new molecular markers to help in the selection of cacao genotypes is a potential strategy (Guimarães and Costa, 2002) to speed cacao programs. In the last few years breeders around the world have used genomic research to identify resistance genes to WBD. Expressed sequence tags (ESTs) elicited with the pathogen *M. perniciosa* (Lanaud et al., 2006; Gesteira et al., 2007; Leal Jr. et al., 2007) and with non-biotic agents (Jones et al., 2002; Verica et al., 2004) have been intensely obtained and analyzed. These ESTs constitute a potential source of new genes related to witches' broom resistance in cacao, as well as of non-neutral polymorphism (e.g., polymorphism occurring in the coding sequence or in its vicinity and directly related to its function).

Among the different kinds of markers useful for developing MAS, single nucleotide polymorphisms (SNPs) have recently captured the attention of the scientific community. They constitute the most common type of sequence differences between alleles and may be used in the detection of associations between phenotype and allelic forms of a gene, especially for common diseases that have multifactorial genetics (Guimarães and Costa, 2002). Due to their potential for higher polymorphisms, SNPs have been developed with success in several plant species such as corn (Ching et al., 2002), *Arabidopsis* (Cho et al., 1999), sugarcane (Grivet et al., 2003), and wheat (Somers et al., 2003).

There are several strategies, both experimental and *in silico*, for SNP discovery and mapping. Experimental SNP discovery consists in a number of laborious steps that make this process complex and expensive. *In silico* discovery has been proposed as an alternative method that benefits and allows the exploitation of private and public data sets for search of potential SNPs (Guimarães and Costa, 2002; Kota et al., 2003). The *in silico* techniques allow

the detection of a virtually limitless number of molecular markers covering the whole genome of an organism (Cho et al., 1999; Picoult-Newberg et al., 1999). Herein, we report the first detection and analysis of SNPs in ESTs from the cacao-*Moniliophthora perniciosa* interaction using *in silico* detection. These SNPs may be considered as potential markers for subsequent use in MAS breeding programs of cacao for witches' broom resistance.

One hundred and fifty-three genes potentially related to plant-pathogen interaction were selected from cacao-*M. perniciosa* interaction libraries available in our laboratory (UESC and CEPLAC projects granted by FAPESB and CNPq; Gesteira et al., 2007) (Table 1) and re-sequenced twice in forward direction on ABI3100 using the primer M13-F 5'-TAAAACGACGGCCAGT-3' and the Big DyEnamic Terminator V. 3.0 kit according to manufacturer instructions (Applied Biosystems/http://www3.appliedbiosystems.com/AB-Home/index.htm) to eliminate sequencing artifacts. Each sequence was analyzed as follows: i) the largest sequenced stretch with Phred quality 10 was extracted using a Perl script (Ewing et al., 1998), ii) the plasmid vector sequence with cross-match (-minmatch 20, -minscore 5) was removed, and iii) the poly(A) tail was removed. Potential sequencing errors were eliminated by comparing the several runs and electropherograms obtained for each individual gene.

Table 1. Characteristics of the cDNA libraries used in this putative single nucleotide polymorphism (SNP) identification.

Cultivar	Phenotype	Tissue	Library type	No. of sequences	Unigene	Contig	Reference
TSH1188	R/inoculated	Meristem	Full-length	3613	1520	199	Gesteira et al., 2007
	R/inoculated	Pod	Full-length	1274	675	191	CEPLAC project[a]
Amelonado/P7B/R10/ Spec54/UF221/Sic5	R and S induced	Leaf and seed	Full-length	4455	1380	nd	Jones et al., 2002
Forastero	Without treatment	Leaf	SSH	1256	nd	nd	Verica et al., 2004
Catongo	S/inoculated	Meristem	Full-length	3271	1065	142	Gesteira et al., 2007

Expressed sequence tags from the TSH1188 libraries were used as reference in the multiple sequence alignments. nd: non-determined. R = resistant; S = susceptible; induced: with salicylic acid, ethylene and jasmonate; inoculated: with *M. perniciosa* spores. [a]Project conducted at CEPLAC and funded by FAPESB and CNPq.

The 153 genes from TSH1188 (resistant genotype) and Catongo (susceptible genotype) (Table 1; Gesteira et al., 2007; CEPLAC research projects supported by FAPESB and CNPq) were then compared with publicly available libraries (Table 1; Jones et al., 2002; Verica et al., 2004) by BLASTX and TBLASTX, adopting an expected value $\leq 1.10^{-4}$ as being significant. TSH1188 sequences showing similarity and having a size larger than 170 bp were used as reference sequences in the alignment process with similar sequences from the other cDNA libraries using the CLUSTAL W program (http://align.genome.jp/; Thompson et al., 1994). The results of the alignment were processed with the CLOURE-D program (Kohli and Bachhawat, 2003) allowing the identification of putative SNPs.

The identification of SNPs by sequence alignment is a very quick and interesting method for comparing sequences from different genotypes. On the other hand, the method of sequence alignment is dependent on the availability of sequences in the database and does not identify mistakes due to sequencing. However, in our case, those difficulties were eliminated by the existence of a cacao library of sequencing data obtained in several previous studies (Jones et al., 2002; Gesteira et al., 2007; Verica et al., 2004) and current work carried out by the CEPLAC re-

search group, and by the adopted strategy of re-sequencing the selected gene for SNPs identification. After trimming, sequence analysis and alignment, processed with the program CLOURE D, only 44 (28.75%) sequences showed potential for SNP identification. The 44 sequences showed a total of 164 substitutions between the genotypes analyzed. To prevent the detection of false SNPs, only alignments containing four or more overlapping sequences were analyzed for polymorphism. Still, to avoid false SNPs and improve the quality of data, only putative SNPs that had a redundancy of two or more per alignment were chosen. This strategy has been adopted in other studies involving the *in silico* detection of SNPs (Brumfield et al., 2003; Batley et al., 2003). Thus, of 44 sequences only 27 (61.36%) were used to detect SNPs, and only 11 sequences (25%) showed the criteria described above, giving a total of 71 putative SNPs (Table 2).

Nineteen putative SNPs were found in two genes of *translationally controlled tumor protein homolog* (TCTP), 15 putative SNPs in the *putative senescence-associated protein* gene, three putative SNPs in the *pentameric polyubiquitin* gene, and one putative SNP in the *ankyrin repeat protein family* gene (Table 2). All these genes are involved in abiotic stimuli and development. Sixteen putative SNPs were identified in the *proline-rich protein* gene, which belongs to the miscellaneous functional class. Ten putative SNPs were identified in *non-cell-autonomous heat shock cognate protein 70* gene, and one putative SNP in the *CuZn-superoxide dismutase* gene; these two genes are involved in signal transduction and post-translational regulation. Three putative SNPs were identified in putative *calmodulin-binding protein* gene, which is involved in secondary metabolism and hormone metabolism. In the *superoxide dismutase* and *thioredoxin H-type (TRX-H)* genes, involved in primary metabolism, there were two and one putative SNP identified, respectively (Table 2).

Because some of these genes belong to multigene families, it may be possible that some of the putative SNPs identified may represent differences between the members of a gene family, rather than differences between alleles. Most of the putative SNPs were found in genes involved in abiotic stimuli and development (54%), miscellaneous (23%), signal transduction and post-translational regulation (15%), secondary metabolism and hormone metabolism (4%), and primary metabolism (4%) (Figure 1).
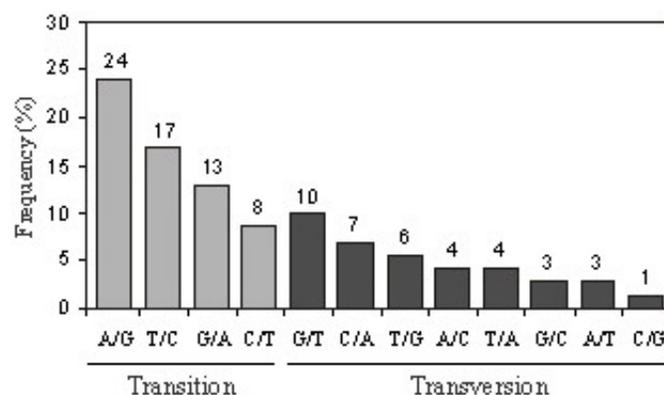

Figure 1. Distribution of putative single nucleotide polymorphisms (SNPs) into functional classes. The 14 broad categories used for classification during the semi-automatic annotation are indicated as well as the number of putative SNPs in each category.

Table 2. Function of 71 putative single nucleotide polymorphisms (SNPs) identified in this study.

Gene	Position[a]	SNP[b]	Number of sequences with SNP[c]	Functional class[d]
Ankyrin repeat family protein	205	G/A	10	XII.B
Non-cell-autonomous	216	T/A	3	X
heat shock cognate protein 70	219	C/T	3	
	230	A/G	3	
	233	T/C	3	
	236	A/G	3	
	249	A/G	3	
	252	A/G	3	
	253	T/C	3	
	255	G/T	3	
	264	C/T	3	
Pentameric polyubiquitin	241	A/G	7	XII.B
	250	T/C	3	
	265	C/T	3	
Proline-rich protein	97	T/G	10	XII
	96	C/A	10	
	95	C/A	10	
	94	G/A	2	
	82	A/T	12	
	85	G/A	15	
	115	A/G	13	
	121	T/C	4	
	124	A/G	10	
	134	C/A	15	
	135	C/A	15	
	136	A/G	15	
	139	G/A	13	
	142	A/G	5	
	145	A/G	28	
	154	G/A	22	
Putative CuZn-superoxide dismutase	111	T/C	2	X
Putative calmodulin-binding protein	235	T/C	3	VI
	242	G/A	2	
	246	A/G	2	
Putative senescence-associated protein	187	A/G	8	XII.B
	188	A/C	8	
	190	C/T	8	
	191	A/G	8	
	192	G/T	2	
	193	T/C	8	
	195	G/A	8	
	197	A/G	2	
	198	T/G	6	
	199	C/T	6	
	202	C/T	6	
	203	G/T	6	
	206	G/A	6	
	207	A/C	6	
	208	G/C	3	

Table 2. Continued.

Gene	Position[a]	SNP[b]	Number of sequences with SNP[c]	Functional class[d]
Superoxide dismutase	132	C/G	3	V
	162	T/G	2	
Thioredoxin H-type (TRX-H)	251	A/G	3	V
Translationally controlled	189	G/T	3	XII.B
tumor protein homolog (TCTP)	190	T/A	3	
	192	T/C	3	
	195	T/C	3	
	203	T/G	3	
	217	T/A	3	
	220	T/C	3	
	224	G/A	3	
	226	T/C	3	
Translationally controlled	349	G/T	3	XII.B
tumor protein homolog (TCTP)	353	A/G	3	
	355	G/C	3	
	362	A/G	3	
	363	T/C	3	
	366	A/T	3	
	367	G/T	3	
	368	G/T	3	
	372	C/A	3	
	373	A/C	3	

[a]From the 5' end of the sequence. [b]Total number of SNPs = 71. [c]Number of sequences showing a base replacement in each alignment. [d]According to Gesteira et al. (2007) and Figure 3 of this study.

Considering the criteria of redundancy (i.e., the frequency of a given SNP in an alignment), of the 71 putative SNPs obtained, 35 had a redundancy of 3. In the same way, 7 putative SNPs had a redundancy of 2, 6 a redundancy of 6, 6 a redundancy of 8, and 5, 4 and 2 a redundancy of 10, 15 and 13, respectively. Putative SNPs with redundancy 4, 5, 7, 12, 22, and 28 were observed only once (Table 2). Transition and transversion SNP types were encountered in all the libraries analyzed, although the transition type appeared with a higher frequency than did the transversion one (62% of transitions, 38% of transversions; Figure 2).

SNP identification has also been observed in other crops (Picoult-Newberg et al., 1999). As a transversion generates a distortion of DNA molecule width and, consequently, may generate serious functional damage in the cell, it has been suggested that this kind of mutation is either rarer or more difficult to be corrected by the DNA system repair (Calcagnotto, 2001). Of the putative SNPs of the transition type (Figure 2), the most represented corresponded to the substitution of guanine for adenine (A/G; 24%), followed by substitution of cytosine for thymine (T/C; 17%) and adenine for guanine (G/A; 13%). Of the putative SNPs of the transversion type substitution, 10% were substitutions of thymine for guanine (G/T) and 7% of adenine for cytosine (C/A). The other 29% corresponded to G/A transitions and T/G, A/C, T/A, G/C, A/T, and C/G transversions (Figure 2).
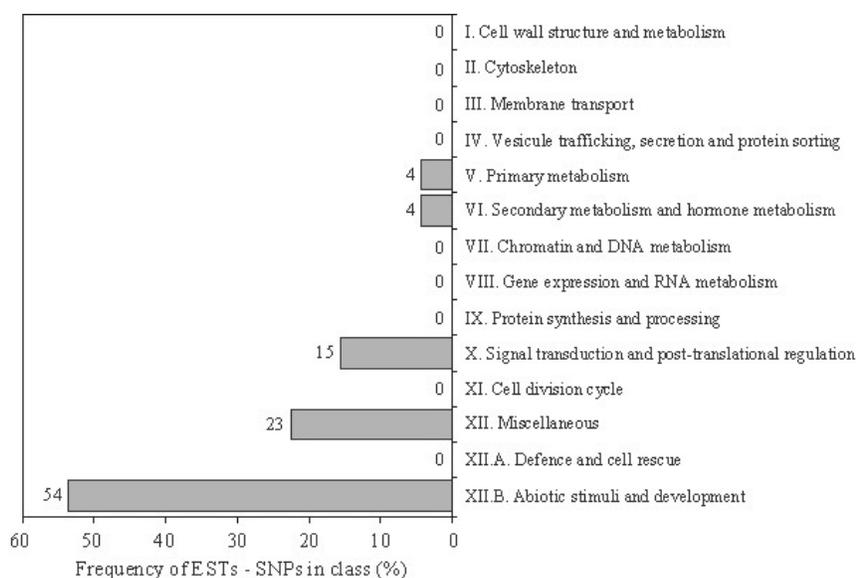
Figure 2. Distribution of different substitution types identified in the cacao-*Moniliophthora perniciosa* expressed sequence tag (EST) sequences. SNPs = single nucleotide polymorphisms.

The position of the 71 SNPs was determined with respect to the open reading frame (ORF) of the corresponding ESTs, and categorized in non-synonymous (leading to amino acid alteration) or synonymous (without amino acid alteration) mu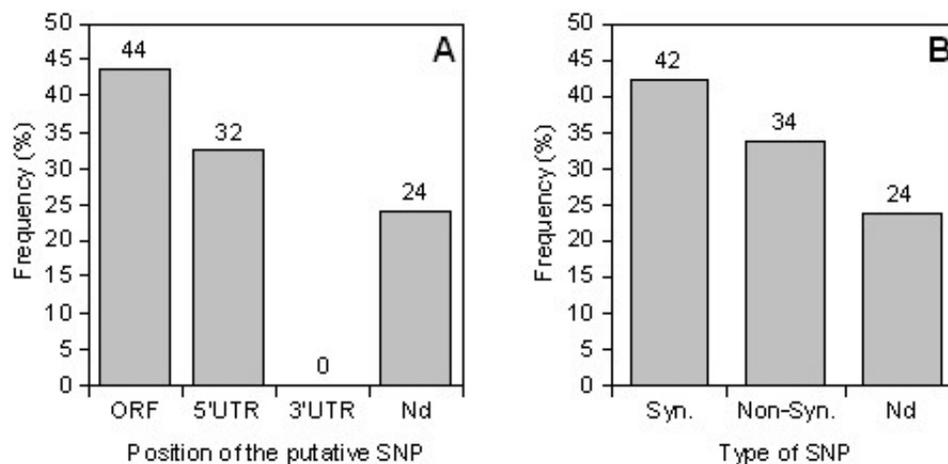tation by scanning sequences for ORFs using the ORF Finder script (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) (Figure 3).

Figure 3. Frequency of single nucleotide polymorphisms (SNPs) in relation to position in ORF and protein synthesis. **A.** Frequency of SNPs in the different sequence parts. Nd: not determined; ORF: open reading frame; UTR: untranslated region. **B.** Frequency of synonymous (Syn.) and non-synonymous (Non-Syn.) SNPs.

A high frequency of SNPs was found in the ORF (44%) and in the 5'UTR (untranslated region) (32%) while no SNPs were observed in the 3'UTR. Twenty-four percent of the SNPs were not located with precision in relationship to the ORF (Figure 3A). The condition of forward sequencing may explain the absence of SNPs in the 3'UTR; the 5'UTR was sequenced, and the 3'UTR was, nevertheless, either partially sequenced or not sequenced depending on the length of the ORF, making it difficult to detect SNPs in this region. The presence of a great amount of SNPs in ORFs was unexpected and contradicted previous works showing that the UTRs have a higher frequency of SNPs than does the rest of the genome (Zhu et al., 2001; Ching et al., 2002). The presence of SNPs in the transcripts of genes suggests that they could play a role in gene expression or function.

In our study, the frequency of polymorphisms in the cacao ESTs is one putative SNP per 223 bp in non-coding regions, and one putative SNP per 71 bp in the coding regions. In maize, the frequency of polymorphisms in the US elite inbreeding lines is one SNP per 31 bp in non-coding regions, and one SNP per 124 bp in the coding regions (Ching et al., 2002). In soybean, a study of sequence diversity in 22 diverse genotypes found 1.64 SNPs per kb in coding regions, and 4.85 SNPs per kb in non-coding regions (Zhu et al., 2001). As observed in Figure 3B, we found more frequent synonymous SNPs (42%) than non-synonymous SNPs (34%). SNPs, when synonymous, can change the structure and stability of the messenger RNA, whereas non-synonymous can change the structure and function of the protein and consequently affect the amount of protein produced and presence of deleterious mutations. Furthermore, SNPs can promote changes in the pattern of gene expression (such as changes in promoter sequences), generation or deletion of termination codons and amendment of the codons for initiation of translation (Guimarães and Costa, 2002). The higher frequency of synonymous SNPs compared to non-synonymous SNPs found herein, indicates a possible selection against deleterious mutations (Zhu et al., 2003). It has been reported in other studies that synonymous SNPs occur more often than non-synomymous ones (Calcagnotto, 2001).

Taking into account the initial number of sequences (153) selected for polymorphism analysis, the numbers of putative SNPs (71; 0.46 SNP per sequence) detected in our study are the same or higher than observed in other studies. For example, in barley, a total of 3069 candidate SNPs were obtained from the analysis of detection algorithm (SNiPpER) in 8171 EST sequences (0.37 SNP per sequence; Kota et al., 2003). In *Lycopersicon esculentum*, it was possible to obtain 101 potential SNPs from 1245 contigs, using *in silico* detection (0.08 SNP per sequence; Yang et al., 2004). Several methods have been used for SNP identification and, among them, discovering SNPs through data sequencing seems to be the most promising, efficient and rapid method (Gu et al., 1998). Moreover, the identification of SNPs from ESTs appears to be a potent approach to investigate markers in coding sequences and to relate markers to cellular function. Here, as prerequisite for use in cacao breeding, we report the first detection and analysis of SNPs from cacao-*Moniliophthora perniciosa* interaction ESTs using bioinformatics.

ACKNOWLEDGMENTS

REFERENCES

Agrios GN (2005). Genetics of Plant Disease. In: Plant Pathology. 6th edn. Academic Press, San Diego, 489-498.

Aime MC and Phillips-Mora W (2005). The causal agents of witches' broom and frosty pod rot of cacao (chocolate, *Theobroma cacao*) form a new lineage of Marasmiaceae. *Mycologia* 97: 1012-1022.

Batley J, Barker G, O'Sullivan H, Edwards KJ, et al. (2003). Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiol.* 132: 84-91.

Brumfield RT, Beerli P, Nickerson DA and Edwards SV (2003). The utility of single nucleotide polymorphisms in inferences of population history. *Trends Ecol. Evol.* 18: 249-256.

Calcagnotto D (2001). Taxas de Evolução e o Relógio Molecular. In: Biologia Molecular e Evolução (Matioli SR, ed.). Ed. Holos, Ribeirão Preto, 51-63.

Ching A, Caldwell KS, Jung M, Dolan M, et al. (2002). SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC Genet.* 3: 19.

Cho RJ, Mindrinos M, Richards DR, Sapolsky RJ, et al. (1999). Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. *Nat. Genet.* 23: 203-207.

Ewing B, Hillier L, Wendl MC and Green P (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8: 175-185.

Gesteira AS, Micheli F, Carels N, Da Silva AC, et al. (2007). Comparative analysis of expressed genes from cacao meristems infected by *Moniliophthora perniciosa*. *Ann. Bot.* 100: 129-140.

Grivet L, Glaszmann JC, Vincentz M, da Silva F, et al. (2003). ESTs as a source for sequence polymorphism discovery in sugarcane: example of the Adh genes. *Theor. Appl. Genet.* 106: 190-197.

Gu W, Aguirre GD and Ray K (1998). Detection of single-nucleotide polymorphism. *Biotechniques* 24: 836-837.

Guimarães PEM and Costa MCR (2002). SNPs: sutis diferenças em um código. *Biotecnol. Cienc. Desenvolv.* 26: 24-27.

Jones PG, Allaway D, Gilmour DM, Harris C, et al. (2002). Gene discovery and microarray analysis of cacao (*Theobroma cacao* L.) varieties. *Planta* 216: 255-264.

Kohli DK and Bachhawat AK (2003). CLOURE: Clustal Output Reformatter, a program for reformatting ClustalX/ClustalW outputs for SNP analysis and molecular systematics. *Nucleic Acids Res.* 31: 3501-3502.

Kota R, Rudd S, Facius A, Kolesov G, et al. (2003). Snipping polymorphisms from large EST collections in barley (*Hordeum vulgare* L.). *Mol. Genet. Genomics* 270: 24-33.

Lanaud C, Fouet O, Gramacho K, Argout X, et al. (2006). A Large EST Resource for *Theobroma cacao* Including cDNAs Isolated from Various Organs and Under Various Biotic and Abiotic Stresses. In: Proceedings of the 15th International Cocoa Research Conference, San Jose, Costa Rica, 185-191.

Leal GA Jr, Albuquerque PSB and Figueira A (2007). Genes differentially expressed in *Theobroma cacao* associated with resistance to witches' broom disease caused by *Crinipellis perniciosa*. *Mol. Plant Pathol.* 8: 279-292.

Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, et al. (1999). Mining SNPs from EST databases. *Genome Res.* 9: 167-174.

Purdy LH and Schmidt RA (1996). Status of cacao witches' broom: biology, epidemiology, and management. *Annu. Rev. Phytopathol.* 34: 573-594.

Somers DJ, Kirkpatrick R, Moniwa M and Walsh A (2003). Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. *Genome* 46: 431-437.

Tanksley SD, Miller J, Paterson A and Bernatsky R (1988). Molecular Mapping of Plant Chromosomes. In: Chromosome Structure and Function (Gustafson JP and Appels R, eds.). Plenum Press, New York, 157-173.

Thompson JD, Higgins DG and Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22: 4673-4680.

Verica JA, Maximova SN, Strem MD, Carlson JE, et al. (2004). Isolation of ESTs from cacao (*Theobroma cacao* L.) leaves treated with inducers of the defense response. *Plant Cell Rep.* 23: 404-413.

Yang W, Bai X, Kabelka E, Eaton C, et al. (2004). Discovery of single nucleotide polymorphisms in *Lycopersicon*

esculentum by computer aided analysis of expressed sequence. *Mol. Breed.* 14: 21-34.

Zhu YL, Hyatt S, Quigley C, Song QJ, et al. (2001). Single Nucleotide Polymorphisms (SNPs) in Soybean Genes, cDNAs, and Random Genomic Sequence. In: Plant and Animal Genome IX Conference, January 13-17, 2001, San Diego.

Zhu YL, Song QJ, Hyten DL, Van Tassell CP, et al. (2003). Single-nucleotide polymorphisms in soybean. *Genetics* 163: 1123-1134.