Genetics and Molecular Biology, 30, 3 (suppl), 1009-1018 (2007)

www.sbg.org.br

Research Article

Frequency and distribution of microsatellites from ESTs of citrus

Darío Abel Palmieri[1*], Valdenice Moreira Novelli[1*], Marinês Bastianel[1*], Mariângela Cristofani-Yaly[1*], Gustavo Astúa-Monge[1], Eduardo Fermino Carlos[1], Antonio Carlos de Oliveira[2] and Marcos Antonio Machado[1]

[1]*Centro APTA Citros Sylvio Moreira, Instituto Agronômico de Campinas, Cordeirópolis, SP, Brazil.*
[2]*Departamento de Ciências Naturais, Universidade Estadual Sudoeste da Bahia, Campus Vitória da Conquista, Vitória da Conquista, BA, Brazil.*

Abstract

Nearly 65,000 citrus EST (Expressed Sequence Tags) have been investigated using the CitEST project database. Microsatellites were investigated in the unigene sequences from *Citrus* spp. and *Poncirus trifoliata*. From these sequences, approximately 35% of the non-redundant ESTs contained SSRs. The frequencies of different SSR motifs were similar between *Citrus* spp and trifoliate orange. In general, mononucleotide repeats appeared to be the most abundant SSRs in the CitEST database, but we also identify di-, tri-, tetra-, penta- and hexanucleotide repeats. The AG/CT and AAG/CTT were the most common dinucleotide and trinucleotide motifs, with frequencies of 54.4% and 25.2%, respectively. Primer sequences flanking SSR motifs were successfully designed and synthesized. After *in silico* polymorphism analysis, a subset of sixty-eight primers was validated in different *Citrus* spp. and *Poncirus trifoliata*. PCR-amplification revealed polymorphism in citrus with all tested primer pairs and showed the potential of these markers for linkage mapping. Our study showed that the CitEST database can be exploited for the development of SSR markers that can amplify *Citrus* spp. and related genus for comparative mapping and other genetic analyses.

Key words: Citrus spp., *Poncirus trifoliata*, ESTs, microsatellites, *in silico* polymorphism.

Received: July 21, 2006; Accepted: February 13, 2007.

Introduction

Microsatellites, or simple sequence repeats (SSR), are arrays of hypervariable short (1-5 bp) repeat motifs that can be found in both coding and non-coding DNA sequence of higher organisms. These single-locus markers are mainly characterized by high frequency, Mendelian inheritance and codominance. During the last decade, microsatellites have proven to be the marker of choice in plant genetics and breeding research, because of their variability, ease of use, accessibility of detection and reproducibility (Zane *et al.*, 2002). There are now many well-known examples of initiatives using microsatellites for different plant species, including *Citrus* sp. (Kijas *et al.*, 1994; Holton *et al.*, 2002, Kantety *et al.*, 2002, Cristofani *et al.*, 2003).

Microsatellites are Polymerase Chain Reaction (PCR) based, requiring previous sequence identification, primer designing for the conserved flanking regions and amplification of the target repeat. Initially, they were expensive to perform, and library enrichment protocols were widely used to reduce investments. However, new sources of microsatellites have been utilized, which are based on large genome sequencing projects. This was initially limited to species where databanks existed, but the increase in available DNA sequence information, particularly ESTs (expressed sequence tags), has provided new opportunities for development of molecular markers for several annual and perennial plant species. Examples are available for *Arabidopsis* (Delseny *et al.*, 1997), grape (Scott *et al.*, 2000), cereals (Kantety *et al.*, 2002), eucalyptus (Ceresini *et al.*, 2005) and others. More recently in citrus, microsatellites were investigated and characterized from public EST-database (Chen *et al.*, 2006; Dong *et al.*, 2006).

Microsatellites based on EST libraries (EST-SSRs) are powerful tools for genetic research in genetic variation, gene tagging and evolution, mapping and analysis of quantitative traits (Cato *et al.*, 2001; Scott, 2001; Holton *et al.*, 2002). In addition, microsatellites can also be used across species (Scott *et al.*, 2000). EST-derived microsatellites have been observed to have high conserved flanking sequences among related species. This characteristic can be

Send corresponcence to Mariângela Cristofani-Yaly. Laboratório de Biotecnologia, Centro APTA Citros Sylvio Moreira, Instituto Agronômico de Campinas, Rodovia Anhanguera km 158, Caixa Postal 4, 13490-970 Cordeirópolis, SP, Brazil. E-mail: mariangela@centrodecitricultura.br.
*These authors contributed equally to the article.

used to build comparative maps, identify orthologous loci and map genes of known function, such as genes controlling agronomic traits of interest (Kantety *et al.*, 2002; Varshney *et al.*, 2002). Although EST-derived SSRs have been shown to be less polymorphic than those derived from genomic sequences, they have some inherent advantages: quickly obtained by electronic sorting, unbiased in their repeat type, present in gene rich regions of the genome, and normally abundant (Scott, 2001).

Our laboratory, 'Centro APTA Citros Sylvio Moreira,' has developed a large citrus EST (CitEST) database, using libraries that represent different physiological conditions and citrus species. Because of the large number of sequences available from the CitEST project, this database was used to search for hypervariable motifs, such as microsatellites. Although most citrus types exhibit clear morphological variation among them, mainly within genus and species, many agronomical traits are difficult to select by conventional techniques, making assistance by molecular markers highly desirable. A potential reason for that is because most of the wanted traits are apparently quantitatively inherited (Cristofani *et al.*, 2003; Novelli *et al.*, 2006). Therefore, in this study, we mined the CitEST database searching for microsatellites, using an *in silico* approach for marker development, and an *in vivo* validation of candidate polymorphic markers. We developed a bioinformatic tool, named MarkerXplorer, which uses several publicly available software programs to retrieve and characterize microsatellite loci from the database. Different citrus genotypes, including a zygotic progeny from Rangpur lime (*Citrus limonia* Osbeck) *vs.* Swingle citrumelo (*Citrus paradisi* Macf. x *P. trifoliata*), were used to validate a set of markers developed in this study.

Material and Methods

The CitEST database

The `Centro APTA Citros Sylvio Moreira' has a project to create and maintain a databank based on ESTs (CitEST) from different physiological conditions and also from different genera and species of citrus. The total number of sequences available is an constantly changing, but up to the date of this study, 171,430 citrus ESTs from different genotypes such as: *Citrus aurantium* L. - sour orange (CA), *C. sinensis* (L.) Osbeck - Pera sweet orange (CS), *C. latifolia* Tan. - Tahiti lime (LT), *C. reticulata* Blanco - Ponkan mandarin (CR), *C. aurantifolia* (Christm) Swingle - Mexican lime (CG), *C. limonia* Osb. - Rangpur lime (CL) and *P. trifoliata* (L.) Raf. - trifoliata orange (PT) were analyzed (http://biotecnologia.centrodecitricultura.br/).

Sorting sequences through a pipeline

MarkerXplorer pipeline uses several scripts and executable programs such as MISA (MIcroSAtellite identification tool) - (Thiel *et al.*, 2003), Primer3 (Rozen and Skaletsky, 2000), and e-PCR (Schuler, 1997), and Perl scripts to identify repeated sequences, to design specific primer pairs and to evaluate, *a priori*, potential polymorphic markers. To avoid redundancy on further analysis, a clustering analysis with CAP3 software (Huang and Madan, 1999) was previously performed, and the MarkerXplorer pipeline was then used to run from a multi-FASTA formatted file containing 64,726 assembled sequences representing 54,492 kb.

Microsatellite search

The adjusted parameters for our pipeline looked for mononucleotides motifs larger than 10 repeat units and all other repeats (di-, tri-, tetra-, penta- and hexanucleotides) larger than five repeat units. To identify the relative position of a SSR within a given sequence we used the strategy adapted by Ceresini *et al.* (2005) that categorized the microsatellites as initial (I, close to the 5' end), middle (M) or end (E, close to the 3' end). Mononucleotides were excluded from this analysis.

Primer design

One of MarkerXplorer output files was a tab-delimited file with all primer pairs flanking sequences nearby the microsatellite. The length of the amplicons was set to 100-350 bp. Oligonucleotide parameters for Primer3 were set to a length of 18-27 bp with an optimum of 20 bp, a GC content of 20%-80% with an optimum of 50%, a melting temperature (Tm) of 57-63 °C with an optimum of 60 °C, and a primer Tm maximum difference of 1 °C. From a list of nearly 2,000 potential EST-SSR markers, 68 primer pairs were selected randomly and synthesized by IDT (Coralville, IA, USA) and tested. These oligonucleotides were resuspended to 100 µM and tested on all different genotypes described below.

Citrus progeny

For primer validation, plant samples were obtained from the germplasm collection of the 'Centro APTA Citros Sylvio Moreira', Cordeirópolis, SP, Brazil, representing the following citrus genotypes: Pera sweet orange, Murcott tangor (*C. sinensis* x *C. reticulata*), Cravo mandarin (*C. reticulata*), Sunki mandarin (*C. sunki* Hort. ex. Tan.), Trifoliata orange, pummelo (*C. grandis* Osbeck), Rangpur lime, Swingle Citrumelo (*C. paradisi* Macf. x *P. trifoliata*) and a zygotic progeny from Rangpur lime *vs.* Citrumelo Swingle. DNA extraction was performed following the method described by Murray and Thompson (1980).

Amplification and analysis of microsatellite loci

PCRs were performed in a final volume of 25 µL, and the reactions contained 2.5 µL of 10X Buffer (100 mM Tris-HCl pH 8.3; 500 mM KCl; 25 mM $MgCl_2$; 0.01% gelatin) (Invitrogen, São Paulo, Brazil), 200 µM of each dNTP (Invitrogen), 4 pmol of each primer (forward and reverse),

100-150 ng of genomic DNA and 1.0 unit of Taq DNA polymerase (Invitrogen). The amplification was performed using a touchdown program (TD1) which was set up to run 30 cycles of 30 s at 94 °C, 30 s at 65-56 °C (touchdown 0.3 °C every cycle) and 5 s at 72 °C. Amplification products were resolved on 4% agarose gels (TAE 1X), containing 0.5 ng/mL of ethidium bromide.

Functional characterization

Functional annotation of Citrus markers was obtained from GenBank using blastX algorithm against nr database (Altschul *et al.*, 1997) and further classified by gene ontology (Ashburner *et al.*, 2000). GO Terms were extracted from the best homologous hit. The AmiGO term browser was used to find molecular function, cellular component and biological process ontology for these sequences.

Results

Frequency and distribution of EST-SSRs in the CitEST database

Using the MarkerXplorer pipeline, we obtained a detailed analysis of the frequency and distribution of all mono-, di-, tri-, tetra-, penta- and hexanucleotides repeats from the six *Citrus* and one *Poncirus* species. A set of 64,726 clustered sequences, with average length ranging from 586-bp (CL) to 891-bp (CS), were screened and 21,584 sequences (33.3%) containing 27,656 non-redundant SSRs were identified (Table 1). Considering that approximately 54.5 Mb were analyzed, we detected a frequency of at least 1 SSR per 1.97 kb in the expressed fraction of citrus genome. From the number of SSR-containing sequences we observed an average frequency of 22.3% of sequences containing more than 1 SSR, with *C. sinensis* and *P. trifoliata* showing the highest frequencies, 24.0 and 23.4%, respectively (Table 1).

Of those EST-SSRs identified by MarkerXplorer pipeline, 3,030 (10.9%) were represented by compound microsatellites. Despite the differences in the total number of examined sequences among the seven species, frequencies of compound microsatellites were very similar, ranging from 9.7% in *C. aurantium* to 11.6% in *C. sinensis* (Table 1).

The most frequent microsatellite types were mononucleotide repeats in all seven studied species. Among the other repeats, di- and trinucleotides showed the highest frequencies for all species, ranging from 4.9% in *C. limonia* to 22.2% in *C. sinensis* and 2.2% in *C. limonia* to 11.4% in *C. sinensis*, respectively. Tetra-, penta- and hexanucleotide repeats were represented in proportions of 0.3 to 0.9%, 0.05 to 0.2% and 0.15 to 0.3%, respectively (Table 2).

Among the mononucleotide repeats, the motif A/T was the most common (87.3%) followed by C/G (12.7%). The AG/CT and AT/AT motifs were the most common dinucleotide repeats with 54.4% and 22.3%, respectively.

Table 1 - Abundance of EST-SSRs in the seven citrus species from CitEST database.

Species	*C. aurantium* (CA)	*C. sinensis* (CS)	*C. latifolia* (LT)	*C. reticulata* (CR)	*C. aurantifolia* (CG)	*C. limonia* (CL)	*P. trifoliata* (PT)	Total
Total number of examined sequences (CitEST)	4,172 (5,950[1])	25,180 (86,472)	3,421 (5,484)	15,470 (40,545)	3,633 (6,621)	2,722 (4,185)	10,128 (24,412)	**64,726 (171,430)**
Total length of examined sequences (kb)	3,368	22,442	2,670	13,316	2,802	1,597	8,297	**54,492**
Total number of identified SSRs	1,673	9,482	1,661	5,822	1,696	2,147	5,175	**27,656**
Number of SSR-containing sequences	1,352 (32.4%)	7,220 (28.7%)	1,350 (39.5%)	4,568 (29.5%)	1,324 (36.4%)	1,762 (64.7%)	4,008 (39.6%)	**21,584 (33.3%)**
Number of sequences containing more than 1 SSR	268 (19.8%[2])	1,730 (24.0%)	257 (19.0%)	988 (21.6%)	300 (22.6%)	329 (18.7%)	938 (23.4%)	**4,810 (22.3%)**
Number of compound SSRs	163 (9.7%[3])	1,101 (11.6%)	171(10.3%)	608 (10.4%)	193 (11.4%)	220 (10.2%)	574 (11.1%)	**3,030 (10.9%)**

[1]Total number of sequences before clustering analysis.
[2]Frequency of sequences containing more than 1 SSR over number of SSR-containing sequences.
[3]Frequency of compound SSRs over total number of identified SSRs.

Table 2 - Frequency distribution (%) of EST-SSRs based on motif size for each species. Numbers within parentheses represent absolute number of microsatellites.

	CA	CS	LT	CR	CG	CL	PT	Total per class
Mononucleotide	76.2 (1,275)	65.1 (6,175)	80.8 (1,342)	69.2 (4,030)	73.2 (1,242)	92.4 (1,985)	79.8 (4,132)	20,181
Dinucleotide	15.5 (260)	22.2 (2,103)	13.3 (221)	19.0 (1,109)	17.2 (291)	4.9 (106)	12.6 (651)	4,741
Trinucleotide	7.3 (122)	11.4 (1,085)	5.1 (84)	10.8 (626)	8.6 (146)	2.2 (47)	6.9 (356)	2,466
Tetranucleotide	0.6 (10)	0.9 (83)	0.5 (09)	0.7 (41)	0.6 (10)	0.05 (01)	0.4 (23)	177
Pentanucleotide	0.1 (01)	0.15 (12)	0.1 (02)	0.05 (02)	0.2 (03)	0.3 (06)	0.1 (05)	31
Hexanucleotide	0.3 (05)	0.25 (24)	0.2 (03)	0.25 (14)	0.2 (04)	0.15 (02)	0.2 (08)	60

For tri- and tetranucleotides, the motifs AAG/CTT reached 25.2%, AAT/ATT 18.4%, AAAT/ATTT 33.0%, and AAAG/CTTT 22.5%, being the most frequent repeats. The remaining penta- and hexanucleotide repeats were represented by AAAAT/ATTTT (19.2%) and AACTC/AGTTG (23.1%), AACACC/GGTTGT and ACGAGG/CCTGCT (8.5%), respectively (Figure 1).

To obtain a more detailed analysis of the SSR structure, we used the categorization proposed by Weber (1990) in three classes: pure or perfect repeats, *e.g.*, $(AT)_n$ or $(CTG)_n$; imperfect repeats, *e.g.*, $(TG)_n(N)_x(TG)_m$ or $(GGC)_n(N)_x(GGC)_m$, and compound repeats, *e.g.*, $(GT)_n(AT)_m$, $(ATC)_n(GCG)_m$ or $(CG)_n(AAT)_m$. For the seven citrus species, we observed a proportion of 89.1% of perfect, 7.2% of compound and 3.7% of imperfect repeats. When considering only perfect repeats, we observed variations in the number of repeat units per microsatellite type and species. The maximal length of the six SSR types ranged from 77 (*C. latifolia*) to 148 (*P. trifoliata*) units for mononucleotides; 20 (*C. limonia*) to 34 (*C. sinensis*) units for dinucleotides; 11 (*C. aurantium*) to 28 (*C. sinensis*) units for trinucleotides; six (*C. limonia*) to (11) units for tetranucleotides; five (*C. reticulata*) to nine (*C. sinensis*) units for pentanucleotides; and six (*C. aurantium*, *C. limonia* and *P. trifoliata*) to 13 (*C. reticulata*) units for hexanucleotides.

We also examined and labeled the relative position of di-, tri-, tetra-, penta and hexanucleotides repeats identified by MarkerXplorer from the seven species. From this analysis we identified that about 58.0% of all microsatellites are localized closer to the 5' end (I), while 23.0% are positioned near the 3' end (E) and 19.0% are distributed along the middle region (M). Among the seven citrus species screened here, *C. limonia* and *C. reticulata* showed the most discrepant values, with 39.7 and 62.8% of the microsatellites categorized as I, 28.7 and 17.6% as M and, 36.6 and 19.6% as E, respectively.

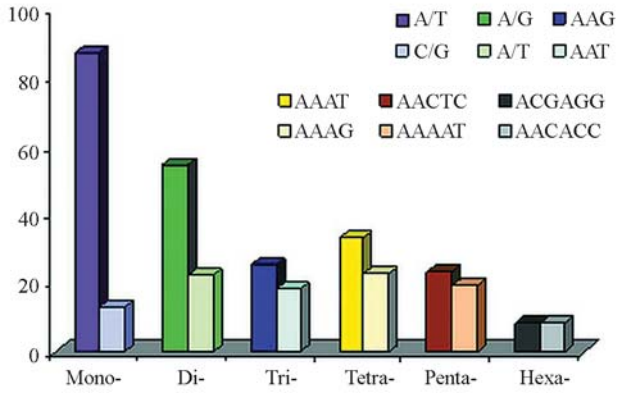

Figure 1 - Frequency of the most common SSR motifs in CitEST database.

According to their potential as genetic markers, Temnykh *et al.* (2001) classified microsatellites of different sizes and types into two categories or groups: those larger than 20 repeated units (Class I); and those equal to or bigger than 12 but smaller than 20 repeated units (Class II). In this study, 31.0% of the EST-SSRs identified were classified as Class I while 69.0% were Class II. The sweet orange

Table 3 - Frequency distribution (%) of the different SSR types (perfect repeats only) according to their potential as genetic markers. Definitions of these classes followed the pattern reported by Temnykh *et al.* (2001).

	CA class		CS class		LT class		CR class		CG class		CL class		PT class	
	I	II	I	II	I	II	I	II	I	II	I	II	I	II
Mononucleotide	0.82	0.75	0.60	0.66	0.88	0.81	0.66	0.69	0.76	0.74	0.95	0.94	0.81	0.82
Dinucleotide	0.05	0.11	0.18	0.17	0.04	0.10	0.16	0.14	0.11	0.13	0.02	0.03	0.08	0.09
Trinucleotide	0.08	0.14	0.15	0.18	0.05	0.08	0.13	0.17	0.09	0.14	0.02	0.03	0.09	0.10
Tetranucleotide	0.03		0.04		0.02		0.04		0.02		0.01		0.02	
Pentanucleotide	0.00		0.01		0.00		0.00		0.01		0.00		0.00	
Hexanucleotide	0.01		0.01		0.01		0.01		0.01		0.00		0.01	

showed the highest number of di- and trinucleotides in both classes (Table 3).

Marker development and *in silico* polymorphism detection

A list of 1,918 primer pairs were obtained from the MarkerXplorer pipeline. From these, 758 showed *in silico* polymorphism ranging from 2 to 10-bp when analyzing sequences deposited into the CitEST database.

To validate microsatellite markers obtained from CitEST database using the MarkerXplorer pipeline, we synthesized 68 primer pairs which were used successfully to amplify PCR products of the expected size in accessions of *Citrus* spp. and the related genus *Poncirus*. Polymorphism was revealed by all primer pairs in the citrus genotypes tested. To analyze the potential of these markers for mapping studies, all functional loci were used in a Rangpur Lime *vs.* Citrumelo Swingle progeny. Twenty-two of these primer pairs (32.0%) were able to reveal polymorphisms with allelic segregation for mapping (Figure 2).

Functional annotation of the EST sequences containing SSRs

Blast searches against the NCBI database were performed for each of the 68 clones that had primer pairs developed. Forty-three sequence (63.2%) matches were identified with several known proteins, while 19 (27.9%) had homology with expressed, hypothetical and unknown proteins from *Arabidopsis thaliana*, *Cicer arietinum* and *Oryza sativa*. Of the remaining sequences, six (8.8%) produced no hits with any known protein (Table 4).

Figure 2 - Amplification for SSR derived from the CitEST database in Rangpur Lime (LC), Swingle Citrumelo (CS) and progeny (h1-h18). Molecular marker of 100 bp (M).

The gene ontology categorization of the 62 sequences that showed some degree of homology revealed that 51 (~82%) of them had a protein match. From these, 40 (82.3%) were homologous to proteins with molecular functions, mainly binding (DNA, RNA and ion), catalytic (protein kinase, lyase and hydrolase) and transporter activities (Figure 3). Forty-three (84.3%) were homologous to proteins involved with the cellular component, mainly nucleus, organelles (chloroplast and mitochondrion) and membrane (Figure 3). Finally, 42 (82.3%) were classified

Table 4 - Description of 68 ESTs-SSRs from the CitEST database.

Marker ID	Primer sequences (5' to 3')	Motif type	Expected product size (bp)*	n_a	Best match from GenBank
CCSMEc1	acgctctctccactatccga ctgcagccgaagatatgtga	$(GAA)_{10}$	215	3	AAO22748 - unknown protein [*A. thaliana*]
CCSMEc2	gcttcttggaatggagcaag cgtttttctgaggtcacggt	$(AT)_{11}$	207	3	CAD56224 - hypothetical protein [*Cicer arietinum*]
CCSMEc3	ccatcatggcttctccagat ttgcatgtgccattgattct	$(TTA)_7$	214	3	AAN03468 - bZIP transcription factor ATB2 [*Glycine max*]
CCSMEc4	cttgctcgagtctacgctcc cttcctcttgcggagtgttc	$(AG)_{14}$	186	1	AAM64478 - ubiquitin-like protein [*A. thaliana*]
CCSMEc5	actgctgttcaccctgttcc gagagctttcgagcctttga	$(CTT)_{10}$	140	4	BAC79192 - unknown protein [*Oryza sativa* (japonica cultivar-group)]
CCSMEc6	gcagcaattctgaaggaagg agtacagcatcctgatcggc	$(TAA)_7$	158	3	AAL85103 - unknown protein [*A. thaliana*]
CCSMEc7	cttggaggaaacagcagagg cgaattggaatcaaaggcat	$(ATC)_8$	155	2	NP_563957 - hydroxyproline-rich glycoprotein family protein [*A. thaliana*]
CCSMEc8	accagagaggctgtgtgctt gtccacgtagtccttgccat	$(GAA)_{11}$	161	2	CAE53883 - aquaporin [*Ricinus communis*]
CCSMEc9	ttcgatagcgctgttgtttg caccatcaccatcacggtag	$(GATGAC)_6$	280	2	AAN41351 - unknown protein [*A. thaliana*]
CCSMEc10	ggtggcgagattatgctgtt tgcagtcccaacaaaaacaa	$(AAC)_7$	272	3	AAO45179 - transcription factor Myb1 [*Malus xiaojinensis*]
CCSMEc11	atctgcagggacaaaaccag tcatcttcactcactcggca	$(GAA)_{10}(n)_{21}$ $(GAA)_7$	213	2	AAM64675 - trihelix DNA-binding protein (GT2) [*A. thaliana*]
CCSMEc12	ggaattcgagttggaggtca accacccatttgcctgataa	$(TC)_{12}$	232	5	O22077 - RuBisCO small subunit, chloroplast precursor
CCSMEc13	atggctttcacagcatctcc tgcatatcctgaagacttttat	$(AT)_{16}$	257	2	BAD27257 - d-limonene synthase [*Citrus unshiu*]

Table 4 (cont.)

Marker ID	Primer sequences (5' to 3')	Motif type	Expected product size (bp)*	n_a	Best match from GenBank
CCSMEc14	gccgatcctctttctctttg aagcacgttatcgggatctg	$(AG)_{15}$ccat $(GGC)_7$	241	4	AAP23944 - leucine-rich repeat protein [x *Citrofortunella mitis*]
CCSMEc15	tgccgttgagttttgattga gactgttgttctgatgccga	$(GAA)_8$	131	3	NP_563819 - expressed protein [*A. thaliana*]
CCSME1	tctttctctttgacaacacaaacc aaccaagtgcagggattgac	$(AG)_{14}$ccat $(GGC)_7$	271	5	AAP23944 - leucine-rich repeat protein [x *Citrofortunella mitis*]
CCSME2	caccaagatcctgaagctga agtacagcatcctgatcggc	$(TAA)_9$	143	3	AAM63195 - unknown [*A. thaliana*]
CCSME3	gccattcacaaaccctctgt aaggaaggggtgttactccg	$(TCC)_7$	279	2	NP_191279 - hypothetical protein [*A. thaliana*]
CCSME4	tagggcgtaacagaattggg gatcggcctcgcataataaa	$(GCA)_8$	170	2	BAB16432 - WRKY transcription factor NtEIG-D48 [*Nicotiana tabacum*]
CCSME5	accatgcaaggagtttccac tttgctttgtgcatggctac	$(CT)_{10}$	185	2	AAC79430 - homeodomain protein [*Malus x domestica*]
CCSME6	attaacgatgatcttggccg tgcagcaaagaaagcaagaa	$(TGA)_8$	247	4	XP_464320 - unknown protein [*Oryza sativa* (japonica cultivar-group)]
CCSME7	aaagccactggggttttagg tcgcttgcctcttcgttaat	$(CGAG)_5$	243	2	AAM63812 - putative SET protein, phospatase 2A inhibitor [*A. thaliana*]
CCSME8	ctcaattccgctagagccac tgaccttggaagccaaaaac	$(AT)_{12}$	173	4	BAD00012 - expansin [*Malus x domestica*]
CCSME9	gcttgacaccgaaaaccatt caatggcttcaaaaaggagc	$(TA)_{12}$	253	2	AAL34292 - unknown protein [*A. thaliana*]
CCSME10	agcaaccaacgacgataagg gttcttggagagtacgggca	$(GCA)_7$	224	2	AAL85084 - unknown protein [*A. thaliana*]
CCSME11	tcttggaccgaattttctcg ttcacaacatctttcgcctg	$(AAAC)_7$	267	3	AAC28763 - unknown protein [*A, thaliana*]
CCSME12	gcctaggctttccttgtcct tccatgcaccatacaccatc	$(ATGT)_5$	130	2	AAD33072 - secretory peroxidase [*N. tabacum*]
CCSME13	tgaagacggtgttaatgcca ggtttagttgccttgccaaa	$(TC)_{14}$	237	2	NP_179865 - homeobox-leucine zipper protein HAT9 [*A. thaliana*]
CCSME14	gagacagcatcagatcagcg gacccatgcaaagatccagt	$(GCA)_8$	256	4	No hits found
CCSME15	aaccgaagatggagggaact acattcatggccacatctca	$(CTT)_8$	206	4	No hits found
CCSME16	caagcttaacttgctaggacaaaa agggaaagaacgagacagca	$(AG)_{11}$	132	4	AAO72532 - aldehyde dehydrogenase 1 precursor [*Lotus corniculatus*]
CCSME17	aatgcgtgggcaataacttc ttcaatatcggcccaaactc	$(GGC)_7$	226	3	NP_197671 - expressed protein [*A. thaliana*]
CCSME18	caggacccagcttgttcaat atttttgcaggacccagatg	$(ATC)_7$	201	3	No hits found
CCSME19	acttatcttgcacccgacga gaggtctcgaagtcacggag	$(ATA)_7$	275	3	AAL60582 - senescence-associated cysteine protease [*Brassica oleracea*]
CCSME20	catctcagactcctgcacca ccctccacccatatcaagaa	$(CTGCTC)_5$	218	5	CAB51544 - RAD23 protein [*Lycopersicon esculentum*]
CCSME21	aggagcagtcttgcttcgag ctcctctctcatcggctctg	$(ATC)_8$	275	2	NP_201087 - peroxisomal protein (PEX14) [*A. thaliana*]
CCSME22	ctatcggcaaaggagcagtc tctctgcaggtaaggttggg	$(ATC)_8$	178	2	NP_201087 - peroxisomal protein (PEX14) [*A. thaliana*]
CCSME23	gccacaatctgtccctgaat cagcacgaaaagcacgaata	$(CTGCTC)_5$	228	4	CAB51544 - RAD23 protein [*L. esculentum*]
CCSME24	gcttcttggaatggagcaag cgtttttctgaggtcacggt	$(AT)_{11}$	207	3	CAD56224 - hypothetical protein [*Cicer arietinum*]
CCSME25	gttgtgtgtccttgggtcct tctgctttggctctccattt	$(TC)_{10}$	274	2	NP_193606 - shaggy-related protein kinase eta / ASK-eta (ASK7) [*A. thaliana*]
CCSME26	gaagaagaagcacgaggacg ccccaaaaataaagcagcaa	$(GT)_{10}$	189	6	AAK52077 - dehydrin COR15 [*Citrus x paradisi*]

Table 4 (cont.)

Marker ID	Primer sequences (5' to 3')	Motif type	Expected product size (bp)*	n_a	Best match from GenBank
CCSME27	gttttgcttgctttgtgtcg caaacccatctaaagcccaa	$(ATAG)_5$	109	3	AAP44415-40S ribosomal protein S9 [*Lactuca sativa*]
CCSME28	aaaaagaacaggagcaggca agaacccacatgcagaaacc	$(CGG)_7$	243	3	NP_565233 - expressed protein [*A. thaliana*]
CCSME29	accagagaggctgtgtgctt gtccacgtagtccttgccat	$(GAA)_{11}$	161	2	AF141900 - putative aquaporin PIP2-2 [*Vitis berlandieri* x *Vitis rupestris*]
CCSME30	aatcaacggaaacaaatccg ctaagctcaccaggctcacc	$(TCT)_8$	271	2	AAO72664 - wheat adenosylhomocysteinase-like protein [*O. sativa* (japonica cultivar-group)]
CCSME31	ggaattcgagttggaggtca accacccatttgcctgataa	$(TC)_{12}$	230	4	AAG49562 - ribulose-1,5-bisphosphate carboxylase/oxygenase small subunit precursor [*Citrus reticulata*]
CCSME32	ggtgcaaccactggaagaat aagagcaaggggtttgtgtg	$(AT)_{10}$	190	3	NP_568458 - expressed protein [*A. thaliana*]
CCSME33	actgtccgatgaagctttgc taaaatccgcttagcccaaa	$(TC)_{12}$	279	3	AAG49562 - ribulose-1,5-bisphosphate carboxylase/oxygenase small subunit precursor [*Citrus reticulata*]
CCSME34	gagcaggccactttctcaac aatagggccccacttgactt	$(GAGT)_5$	242	2	AAL91203 - putative C2H2-type zinc finger protein [*A. thaliana*]
CCSME35	tgattactctgcctccccct tgcacagtcatgtctcctgg	$(GA)_{10}$	253	3	NP_197570 - senescence-associated protein -related [*A. thaliana*]
CCSME36	cgccatttgcttaggcttt gaaccctccagagttccctc	$(AG)_{10}$	220	3	NP_198801-40S ribosomal protein S9 (RPS9C) [*A. thaliana*]
CCSME37	atctcattctccatcgccac ctctccatttttgcctccaa	$(TC)_{11}$	244	2	Q39011 - Shaggy-related protein kinase eta (ASK-eta) [*A. thaliana*]
CCSME38	tgaaggctttttagagccga aatagggccccacttgactt	$(GAGT)_5$	224	2	No hits found
CCSME39	tccctcgtctcaacacaaca gaggctaatgcaagagaccg	$(CT)_{13}$	136	2	CAB81454 - photosystem II protein W-like [*A. thaliana*]
CCSME40	tgttgctgctgctcaaactt tgaaactgaaaacaaaggattcaa	$(TA)_{12}$	106	6	No hits found
CCSME41	ttcatggcagcttgagtttc agtcatggaagccaaaatgg	$(TC)_{12}$	250	4	AAM44082 - putative sorbitol transporter [*Prunus cerasus*]
CCSME42	gcagcaccaacaaagaatca gaacgtacggggtcaaaaga	$(GCC)_7$	163	2	AAM29150 - citrus sucrose transporter 1 [*Citrus sinensis*]
CCSME43	gccgatcctctttctctttg gttggatcccagctctgaag	$(AG)_{13}$	253	5	AAP23944 - leucine-rich repeat protein [x *Citrofortunella mitis*]
CCSME44	ttttggtttcatctgggtcc cctcctgatgatgtcccact	$(TC)_{11}$	129	2	AAK92213 - bZIP transcription factor BZI-2 [*Nicotiana tabacum*]
CCSME45	aagggttgttgctttggttg ttcgctgaagagcatgtcac	$(GAA)_8$	196	4	AAM63312 - unknown [*A. thaliana*]
CCSME46	tgctcttcggtgtcagagtg cagccccaacaaaaacaact	$(AAC)_7$	257	2	AAO45179 - transcription factor Myb1 [*Malus xiaojinensis*]
CCSME47	agagcaacaaaaagtcccga agggttgtggtaggatcacg	$(TCT)_7$	238	3	NP_178346 - expressed protein [*A. thaliana*]
CCSME48	ccatcatggcttctccagat tactccaattgcatgtccca	$(TTA)_7$	222	3	No hits found
CCSME49	aataagcgtatcagcagcagg aatcatgaacgggctgaaac	$(GTTGA)_6$	229	3	NP_195311 - Ras-related GTP-binding protein, putative [*A. thaliana*]
CCSME50	gagttgggattctgctgttga gactgttgttctgatgccga	$(GAA)_7$	140	2	NP_563819 - expressed protein [*A. thaliana*]
CCSME51	tcattagctgattgccagga aatgggaggatgcatggtaa	$(TA)_{12}$	216	4	NP_173866 - polcalcin, putative / calcium-binding pollen allergen, putative [*A. thaliana*]
CCSME52	ctggctcagctctgctcatt tgctgctgcttctgcttcta	$(TC)_{14}$	182	2	NP_178216 - myb family transcription factor [*A. thaliana*]
CCSME53	gatctccctatcatcggcaa ttttgagggctggatggata	$(TAA)_{11}$	187	3	T09875 - late embryogenesis-abundant protein Lea14-A - upland cotton

*Expected product size (bp) estimated from the Primer3 program.
n_a: *in silico* polimorphism (number of alleles).

as involved in biological processes such as cellular metabolism (nucleotide-excision repair, protein biosynthesis, protein amino acid phosphorylation and proteolysis), transcription (mainly regulation), transport (sucrose and carbohydrate), response to stimulus (oxidative stress, water and desiccation) and cell organization and biogenesis (Figure 3).

Discussion

The CitEST project has generated a large set of EST sequences from citrus, an excellent resource for rapid discovery of SSRs. Our study clearly illustrates that ESTs are a useful source of new SSR markers for citrus that are polymorphic and can be transferred between species and related genus. A number of reports have demonstrated that EST databases are a very good source of polymorphic markers for many organisms, including plants (Delseny *et al.*, 1997; Scott *et al.*, 2000; Kantety *et al.*, 2002; Ceresini *et al.*, 2005).

The MarkerXplorer pipeline used here was able to screen the CitEST data for all SSR motifs as well as to identify, *in silico*, a large number of primer pairs with a high potential for germplasm characterization and genetic mapping studies. Bioinformatics approaches are increasingly being used for molecular marker development since the sequences from many genomes are made freely available in public databases (Kantety *et al.*, 2002; Varshney *et al.*, 2002). These sources are mined for SSRs using computational tools thereby eliminating the need for costly, laborious and time-consuming marker development.

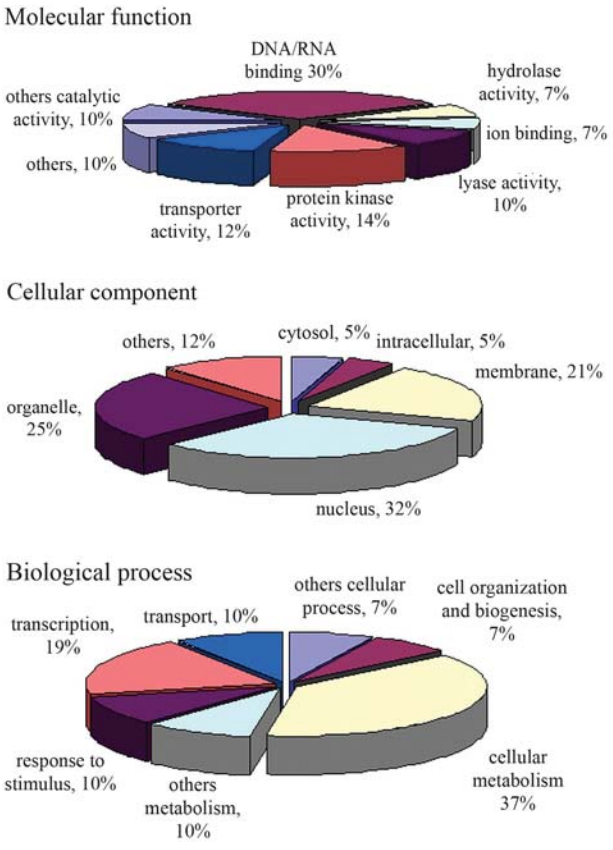

Figure 3 - Citrus EST-SSRs characterization as derived from Gene Ontology categories.

Our results show that SSRs in the CitEST are highly abundant (33.3%) when compared with others crops (Scott *et al.*, 2000; Kantety *et al.*, 2002). In other citrus databases, a frequency of 10.6% of EST sequences with at least one SSR was observed (Chen *et al.*, 2006) and a total of 21.7% in citrus unigene analysis (Dong *et al.*, 2006). Rangpur Lime was the species with the highest number of SSR motifs identified (64.7%). This result should facilitate its use in mapping experiments geared toward molecular breeding. Sweet orange had highest number of sequences analyzed, but this fact was not reflected in a larger number of microsatellites detected.

The dinucleotides AG/CT were the most abundant microsatellites in EST sequences, which is consistent with previous surveys of SSR repeats in annual species, as described by Kantety *et al.*, (2002) in comparative analysis using publicly available EST databases for barley, maize, rice, sorghum and wheat. In fact, this motif was also observed in perennial crops, such as eucalyptus (Ceresini *et al.*, 2005), apple (Newcomb *et al.*, 2006), strawberry (Folta *et al.*, 2005) and citrus (Chen *et al.*, 2006; Dong *et al.*, 2006; Novelli *et al.*, 2006). In general, the adenine-rich repeat motifs are most common in SSRs and, in the CitEST database, these motifs (AAG, AAT, AAAT, AAAG, AAAAT) were also the most abundant (Figure 1). Similar results were reached in others analysis of EST-SSRs in citrus (Chen *et al.*, 2006; Dong *et al.*, 2006) and apple (Newcomb *et al.*, 2006). Additionally, trinucleotide repeats were representative among the different classes and showed the highest level of polymorphism in citrus species (data not show).

The origin and functional role of the microsatellites in expressed sequences are not well understood, but they presumably originate from single or multiple mutational events (Zane *et al.*, 2002). Perfect SSRs were the most frequent type in our study, followed by compound and imperfect types. This suggests a certain degree of sequence stability and conservation in their recent evolution of citrus species. The exploration of perfect SSR in the CitEST database may be a valuable tool to study the evolution of proteins in citrus, since this SSR type is common in many proteins (Katti *et al.*, 2000). Additionally, microsatellites loci with a high number of perfect repeats are usually more polymorphic (Weber, 1990).

The position where a SSR motif occurs in the EST sequence (middle region, 5' and 3' end) can influence the level of polymorphism of a marker (Scott *et al.*, 2000). In our analysis, the 5' end region (I) showed the highest number of repeats, but the level of polymorphism among the species was not estimated in this study.

High-quality amplification products were generated in all EST-SSRs evaluated in citrus species, many of which were polymorphic for at least one species, showing that EST sequences are a useful source of markers for *Citrus* and correlated genera and could therefore be related to agronomic important characters in genetic mapping.

Specific primer pair design from EST sequences on the CitEST database was more efficient when compared to primer design from the genomic libraries previously developed by our group (Novelli *et al.*, 2000, 2006). However, we did not test the hypothesis that microsatellites derived from ESTs are less polymorphic than those derived from genomic libraries as demonstrated by several reports (Scott *et al.*, 2000; Thiel *et al.*, 2003).

The screening of EST-SSRs for polymorphism assessment using an F_1 progeny of the Rangpur Lime *vs.* Swingle Citrumelo cross revealed that 22 of the 68 amplified fragments revealed informative segregation configurations (Figure 2). The markers could be mapped in Rangpur Lime and Swingle Citrumelo according to the pseudo-testcross strategy. This is an encouraging result considering that the number of informative SSR markers is still limited in citrus due to several characteristics related to the biology of these species (Chen *et al.*, 2006; Novelli *et al.*, 2006).

Despite the fact 82% of EST-SSRs markers presented herein showed homology with known gene products, no specific pattern of association to a specific component, process or function was detected. As observed in other works, the majority of transcripts detected represent enzymes of general metabolism (Ceresini *et al.*, 2005; Folta *et al.*, 2005, Newcomb *et al.*, 2006). However, those transcripts related to biological processes such as response to biotic and abiotic stresses can now be readily mapped using existing populations.

In conclusion, our analysis revealed that our CitEST database is a valuable source for rapidly developing new SSR markers, highly transferable for diversity and mapping studies in citrus species and related genera. These data also could be applied in citrus breeding programs such as germplasm characterization, screening of zygotic and nucellar seedlings, and developing markers for marker-assisted selection.

Acknowledgments

We would like to thank all the members of the CitEST project and PADCT/CNPq for financial support. D.A.P. is a postdoctoral fellow funded by CNPq; V.M.N. is FAPESP postdoctoral fellow (Process n° 04/11854-1); M.C.Y. and M.A.M. are recipients of research fellowships from CNPq.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.* (2000) Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium Nat Genet 25:25-29.

Cato SA, Gardner RC, Kent J and Richardson TE (2001) A rapid PCR-based method for genetically mapping ESTs. Theor Appl Genet 102:296-306.

Ceresini PC, Petrarolha Silva CLS, Missio RF, Souza EC, Fischer CN, Guillherme IR, Gregorio I, da Silva EHT, Cicarelli RMB, da Silva MTA, *et al.* (2005) Satellyptus: Analysis and database of microsatellites from ESTs of *Eucalyptus*. Genet Mol Biol 28:589-600.

Chen C, Zhou P, Choi YA, Huang S and Gmitter Jr FG (2006) Mining and characterizing from citrus ESTs. Theor Appl Genet 112:1248-1257.

Cristofani M, Machado MA, Novelli VM, Souza AA and Targon MLPN (2003) Construction of linkage maps of *Poncirus trifoliata* and *Citrus sunki* based on microsatellite markers. Proceedings of the 9th International Society of Citriculture Congress, Orlando, USA, 1:175-178.

Delseny M, Cooke R, Raynal M and Grellet F (1997) The *Arabidopsis thaliana* cDNA sequencing projects. FEBS Lett 405:129-132.

Dong J, Guang-Yan Z and Qi-Bing H (2006) Analysis of microsatellites in citrus unigenes. Acta Genet Sinica 33:345-353.

Folta KM, Staton M, Stewart PJ, Jung S, Bies DH, Jesdurai C and Main D (2005) Expressed sequence tags (ESTs) and simple sequence repeat (SSR) markers from octoploid strawberry (*Fragaria* x *ananassa*). BMC Plant Biol 5:12.

Holton TA, Christopher JT, McClure L, Harker N and Henry RJ (2002) Identification and mapping of polymorphic SSRs markers from expressed gene sequences of barley and wheat. Mol Breeding 9:63-71.

Huang X and Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9:868-877.

Kantety RV, La Rota M, Matthews DE and Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol Biol 48:501-510.

Katti MV, Sami-Subbu R, Ranjekar PK and Gupta VS (2000) Amino acid repeat patterns in protein sequences: Their diversity and structural-functional implications. Protein Sci 9:1203-1209.

Kijas JMH, Fowler JCS, Garbett CA and Thomas MR (1994) Enrichment of microsatellites from the *Citrus* genome using biotinylated oligonucleotide sequences bound to streptavidin-coated magnetic particles. Biotechniques 16:657-662.

Murray MG and Thompson WF (1980) Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res 8:4321-4325.

Newcomb RD, Crowhurst RN, Gleave AP, Rikkerink EHA, Allan AC, Beuning LL, Bowen JH, Gera E, Jamieson KR and Janssen BJ (2006) Analyses of expressed sequence tags from apple. Plant Physiol 141:147-166.

Novelli VM, Cristofani M and Machado MA (2000) Evaluation of microsatellite markers in cultivars of sweet orange (*Citrus sinensis* Osbeck). Acta Horticulturae 535:47-50.

Novelli VM, Cristofani M, Souza AA and Machado MA (2006) Development and characterization of polymorphic simple

sequence repeats (SSRs) in sweet orange (*Citrus sinensis* L. Osbeck). Genet Mol Biol 29:90-96.

Rozen S and Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S and Misener S (eds) Bioinformatics Methods and Protocols in the Series Methods in Molecular Biology. Humana Press, Totowa, pp 365-386.

Schuler GD (1997) Sequence mapping by electronic PCR. Genome Res 7:541-550.

Scott KD (2001) Microsatellite derived from ESTs, and their comparison with those derived by other methods. In: Henry RJ (eds) Plant Genotyping: The DNA Fingerprinting of Plants. CABI Publishing, Oxon, pp 225-237.

Scott KD, Eggler P, Seaton G, Rossetto M, Ablett EM, Lee LS and Henry RJ (2000) Analysis of SSRs derived from grape ESTs. Theor Appl Genet 100:723-726.

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S and McCouch S (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): Frequency, length variation, transposon associations, and genetic marker potential. Genome Res 11:1441-1452.

Thiel T, Michalek W, Varshney RK and Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet 106:411-422.

Varshney RK, Thiel T, Stein N, Langridge P and Graner A (2002) *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell & Mol Biol Letters 7:537-546.

Weber JL (1990) Informativeness of human (dC-dA)n (dG-dT)n polymorphisms. Genomics 7:524-530.

Zane L, Bargelloni L and Patarnello T (2002) Strategies for microsatellite isolation: A review. Mol Ecol 11:1-16.

Internet Resource

AmiGO Term browser, http://www.godatabase.org/cgi-bin/amigo/go.cgi (December 15, 2006).

Associate Editor: Raquel Luciana Boscariol-Camargo