Genetics and Molecular Biology, 24 (1-4), 155-159 (2001)

Preliminary analysis of microsatellite markers derived from sugarcane expressed sequence tags (ESTs)

Jorge A.G. da Silva

Abstract

Expressed sequence tags (ESTs) in the sugarcane (*Saccharum* spp) database (SUCEST) were electronically searched and 402 microsatellites identified. Various dinucleotide and trinucleotide simple sequence repeat (SSR) motifs were found, with these being more frequently observed in ESTs obtained from flower cDNA libraries. PCR primers were designed for 20 of these SSRs and were tested on eight sugarcane genotypes, the sequences of these primers and a list of known sugarcane genes containing SSR motifs being presented in this paper. Polymorphisms were evident both at the cultivar level and between *Saccharum* species. These results show that EST-derived SSRs in *Saccharum* species are useful because they are polymorphic and transferable. The large number of microsatellites that will eventually be available from the SUCEST database (containing 295,000 submitted reads) will have many potential applications in linkage mapping and the planning of crosses.

INTRODUCTION

Polymorphism assays based on variation in the number of short tandemly repeated DNA sequences (microsatellites) have recently been successfully applied to plant breeding programs (Gupta *et al.*, 1996). Once developed, these markers are easy to apply, although the methodology for their development is complex and costly, which limits their application to important crops such as sugarcane (*Saccharum officinarum*) (Scott *et al.*, 2000).

When primers, obtained by the sequence flanking the repeat unit, are used with genomic DNA in PCR reactions they reveal simple sequence repeat (SSR) polymorphisms representing different alleles at that locus (Gupta *et al.*, 1996). In eukaryotic genomes, SSR markers may be identified because interspersed and simple repeats may well overlap regions transcribed by RNA polymerase, including expressed sequence tags (ESTs) in humans; tandem repeat polymorphisms in human genes being more common than generally believed with about 8% of such loci being within the coding sequence and, if polymorphic, resulting in frame shifts (Wren *et al.* 2001). The genetic attributes of microsatellites as DNA markers (co-dominance, high heterozygosity and Mendelian inheritance) coupled with their presence in most, if not all, sugarcane genotypes (Cordeiro *et al.*, 1999) make microsatellites the preferred method for use in the construction of a reliable framework genetic map of sugarcane.

The cost of running microsatellite marker analyses is relatively low, but the cost of developing the markers is high, limiting their application to larger commercial crops such as sugarcane, where enrichment protocols, aiming at increasing the proportion of microsatellite sequences in DNA libraries, have been used with limited success, *e.g.* when library enrichment methods were applied the success rate was only 20% (Cordeiro *et al.*, 1999). The high cost of developing plant microsatellite libraries, coupled with the low level of enrichment peculiar to sugarcane libraries, illustrates the importance of searching for alternative methods of microsatellite development for this crop.

The discovery of SSR markers in ESTs provided the opportunity to develop microsatellites in a simple and direct way, *i.e.* by the electronic searching (data mining) of EST databases. Scott *et al.* (2000) used this new approach in grapes, and described 10 SSRs derived from a grape EST database containing 5,000 ESTs, 2.5% of the total population of cDNA clones were non-redundant microsatellite dinucleotides of seven or more repeats and trinucleotides of five or more repeats, while 10 out of 16 primers yielded products of expected size, all containing polymorphisms, the SSRs being found in both un-translated (5 and 3) as well as in coding sequences.

The sugarcane EST database generated by the SUCEST (sugarcane EST project, sponsored by FAPESP - Fundacao de Amparo a Pesquisa do Estado de Sao Paulo) project is a source of microsatellite-related sequences, allowing the identification of this type of marker and its application for sugarcane breeding. The work presented in this paper describes the discovery of SSRs in the SUCEST database, the first sugarcane microsatellites developed in Brazil, and describes the utilization of 20 SSRs in work involving the genotypes of both commercial and wild sugarcane.

Corresponding author: Soil and Crop Sciences Department, Texas A&M University, 2415 E. Hwy. 83, TX 72596, Weslaco, USA.
E-mail: jadasilva@ag.tamu.edu.

METHODOLOGY

Maize, rice and other plant microsatellite 'words', consisting of short runs of consecutive letters, were used to search the SUCEST nucleotide database (http://sucest.lad.ic.unicamp.br/en/ - FAPESP - Fundacao de Amparo a Pesquisa do Estado de Sao Paulo, Sao Paulo, SP), using the BLASTn software (Scott *et al.*, 2000) for all possible mono, di and trinucleotide repeat patterns. Maize SSRs were obtained from the Missouri Maize Project (University of Missouri - Columbia, ,) and rice SSRs, from the Monsanto Rice Genome Project (Monsanto Organization). Sequences with ten or more mononucleotide repeats, seven or more dinucleotide and five or more trinucleotide repeats were used in this study.

PCR primer pairs were designed using the Primer3 program (Whitehead Institute for Biological Research, www.genome.wi.mit.edu). For each SSR unit present in more than one read the PCR primer pair was designed from the cluster consensus sequence, whereas for those SSR units present in only one read the primer pair was designed from the read sequence.

PCR products were resolved in either high-resolution Agarose 1000 (GIBCO, USA) or polyacrylamide gel at 120 V for four hours, stained with ethidium bromide and visualized under UV light. All the laboratory analyses were conducted at the Copersucar Technology Center molecular biology laboratory (Piracicaba, Sao Paulo, Brazil).

RESULTS

The SUCEST database was used to investigate 85 SSR units (62 common to maize), a total of 402 ESTs (with an *e* value equal or less than -5) being found to be carrying SSRs. At the time the data mining analysis was performed 3-reads composed 11.2% of the total reads and for this reason if an SSR appeared with the same frequency at both the 5 and 3end of sugarcane ESTs one would expect 45 3-reads (out of 402) but only 30 were found, suggesting that SSRs tend to be preferentially present at the 5 end of sugarcane cDNA.

SSRs markers derived from grape ESTs showed that the 3 untranslated region was the most polymorphic at cultivar level (Scott *et al.*, 2000). To test this hypothesis in sugarcane, EST clones sequenced from the 3 end were given priority for designing of PCR primers (see below).

Polymorphism in wild and commercial sugarcane genotypes

The potential utility of SSRs as a molecular profiling technology to aid in sugarcane research and product development remains to be evaluated, although the design of PCR primer pairs for the generation of sugarcane microsatellite markers reported in this work can be used to assess the usefulness of sugarcane EST-SSR markers.

Table I - Polymorphism detected in wild and commercial hybrid sugarcane genotypes.

Marker	SSR	EST	Polymorphism	
			Wild genotype	Commercial genotype
EST-SSR1	(GT)22	SCRUAD1064C07.b	n. a*	n. a.
EST-SSR2	(TC)33	SCJFST1009G12.b	monomorphic	n. a.
EST-SSR3	(GT)32	SCVPRT2083D03.g	n. a.	n. a.
EST-SSR4	(GGC)9	SCQSFL3035C07.b	polymorphic	polymorphic
EST-SSR5	(CA)15	SCRUFL1119B07.b	polymorphic	polymorphic
EST-SSR6	(AAG)21	SCUTFL3071B06.g.	monomorphic	polymorphic
EST-SSR7	(AAC)25	SCJLRZ1018H04.g.	polymorphic	polymorphic
EST-SSR8	(AT)10	SCRLLR1131B11.g.	monomorphic	monomorphic
EST-SSR9	(TTC)20	SCRULB2064F05.b.	monomorphic	monomorphic
EST-SSR10	(AT)47	SCCCST1004E11.g	n. a.	n. a.
EST-SSR11	(AGC)12	SCPRRT3028A10.g	n. a.	n. a.
EST-SSR12	(TTA)138	SCVPHR1092D05.g	n. a.	n. a.
EST-SSR13	(TTTC)15	SCAGLB2047G03.g	n. a.	n. a..
EST-SSR14	(CGG)12	SCEQSB1017A10.g	polymorphic	polymorphic
EST-SSR15	(GA)31	SCCCNR2001E07.g	-	monomorphic
EST-SSR16	(AAAG)11	SCEQHR1082D09.g	monomorphic	monomorphic
EST-SSR17	(ACAA)7	SCBFAD1048D04.b	polymorphic	monomorphic
EST-SSR18	(GCA)15	SCEZAD1082C12.g	monomorphic	polymorphic
EST-SSR19	(TC)50	SCCCCL3080B09.g	polymorphic	polymorphic
EST-SSR20	(TC)50	SCEZHR1048B03.g	polymorphic	polymorphic
EST-SSR21	(AAC)22	SCEZST3151F02.g	-	-
EST-SSR22	(AAT)17	SCRLFL3009C05.b	-	polymorphic

*n.a. = no amplification.

Of the total identified ESTs carrying SSR markers, 20 were chosen for the design of PCR primers (Table I, Appendix I), which were then synthesized and tested in PCR reactions with commercial and wild sugarcane genotypes in order to assess the level of polymorphism and transferability of EST-derived SSRs.

Because the EST-derived microsatellites were within transcribed regions of the DNA there existed the possibility that they would have limited polymorphism, this concern being addressed by using 3 *Saccharum officinarum* (IJ76-418 RED, IN84-3 and KOIKE BLACK) 3 *Saccharum robustum* (IM76-229, IM76-260 and IN84-78) and 2 commercial hybrids (SP80-4966 and SP80-180), as shown on Table II. Figure 1 presents the PCR profile obtained when DNA from these genotypes was amplified using primers flanking the SSR present in a cDNA moiety obtained from a root library.

Out of 20 EST-SSR primer pairs used to study wild sugarcane genotypes, six did not amplify genomic DNA, while of the 14 EST-SSR primer pairs that amplified genomic DNA of two commercial sugarcane hybrids (SP80-4966 and SP80-180), nine showed polymorphism and five yielded monomorphic bands. These two commercial hybrids were chosen for primer screening because they are the parents of a population currently being used for linkage mapping and quantitative trait loci (QTL) tagging, it is planned that the EST-SSR primers showing polymorphism will be mapped in this population.

Most of the primers studied amplified larger fragments than expected, reflecting the possible presence of introns within the genomic DNA sequence, and suggests that the presence of long introns between the sequences homologous to the primers in the genomic DNA may explain the lack of amplification in some of the six pairs that did not amplify.

This sample of EST-SSR markers was compared to a small sample of 13 sugarcane microsatellite markers derived from genomic DNA through the traditional method using a library enrichment technique (Cordeiro *et al.*, 1999). These primers were used in PCR reactions with the same two commercial hybrids shown in Table II. Only three of these 13 primers (23%) detected polymorphism. This preliminary result suggests that EST-derived sugarcane microsatellite markers have the same power to detect polymorphism as genomic microsatellites.

Table II - Wild and Commercial hybrid sugarcane genotypes studied with EST-SSR markers.

Genotype	Status
SP80-4966	Commercial hybrid
SP80-180	Commercial hybrid
IJ76-418 RED	*Saccharum officinarum*
IM76-229	*Saccharum robustum*
IM76-260	*Saccharum robustum*
IN84-3	*Saccharum officinarum*
IN84-78	*Saccharum robustum*
KOIKE BLACK	*Saccharum officinarum*

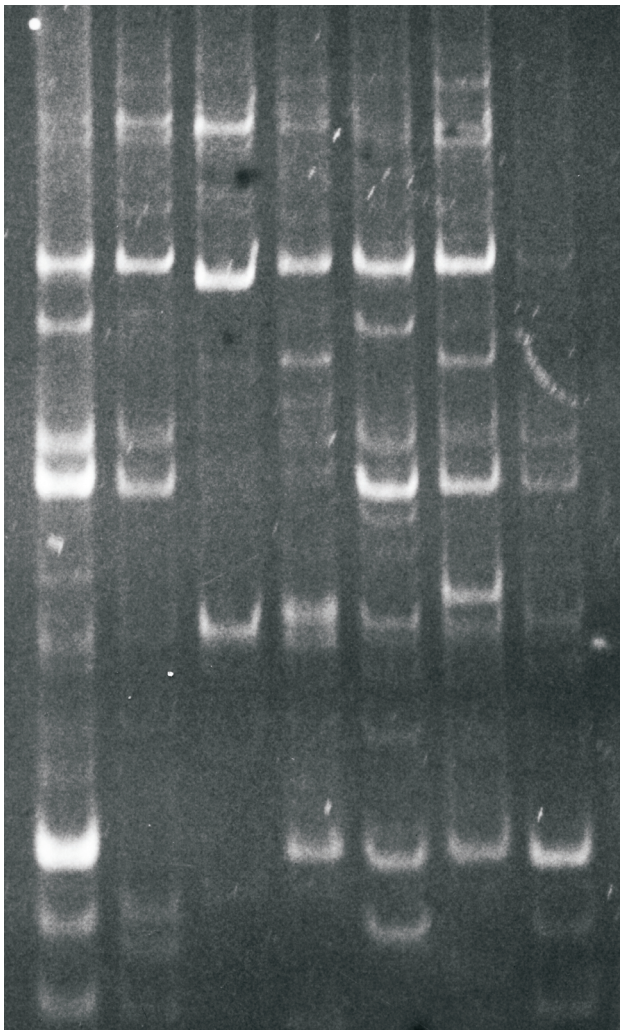

Figure 1 - PCR Product of marker SCJLRZ1018H04.g. Seize Marker: 50 kb ladder; 1 = SP80-4966, 2 = SP80-180, 3 = IJ76-418 Red, 4 = IM76-229, 5 = IM76-260, 6 = IN84-3, 7 = IN84-78.

Frequency of SSR in cDNA libraries

SSRs occur in different expressed sequences of different sugarcane tissues, as can be seem in Table III, this data suggesting that SSRs tend to be more frequent in flower cDNAs and less frequent in etiolated leaves, although this result must be interpreted with caution given the different number of reads obtained from each library.

DISCUSSION

Microsatellite markers are important for sugarcane research because they are PCR markers and are easy to perform, also they are the product of specific primers and are more stable than those generated by random primers such

Table III - Number of EST clones carrying SSRs in the cDNA of different sugarcane tissues.

Type of sugarcane tissue	cDNA libraries	ESTs with SSRs
Plantlets infected with *Gluconobacter diazotroficus*	1	26
Apical Meristem	2	46
Calli[1]	4	16
Flowers (1 to10 cm in diameter)	5	76
Plantlets infected with *Herbaspirillum rubrisubalbicans* (leaf roll)	1	22
Lateral bud	2	21
Leaf roll	2	18
Etiolated leaves	1	6
Root	3	34
Leaf-root transition zone	2	26
Stem Bark	1	24
Seeds	2	51
Internode	2	36
Total		402

[1]Calli were submitted to a light/dark and temperature (4 °C and 37 °C) stress cycle.

as random amplified Polymorphic DNA (RAPD) markers. Another advantage is that, unlike Amplified Length Polymorphism (AFLP) markers, they are locus-specific and are transferable across genotypes within the species, which is an important factor for mapping purposes.

Another possible advantage of using EST-derived SSR markers, besides being easier to develop is that, once mapped, they will be associated with the genes carrying them. In fact, many SSR harboring ESTs show homology to known genes when used for searches with BLASTX. One example is the SCRLFL3009C05.b EST, found in one of the sugarcane flower cDNA libraries, that carries the SSR $(AAT)_{17}$ and codes for a resistance like protein. Another example is the SCSBRZ3121G06.g EST which is homologous to the disease resistance protein AF140722 from *Oryza sativa.* Appendix II shows a list of SSRs carrying EST clones homologous to known genes.

Repetitive elements are quite commonplace in Eukariotic DNA database sequences, even among mRNA sequences, and almost always fall within the untranslated regions. The work on mapping and cloning of several different human hereditary disease genes has led to the discovery of trinucleotide SSR sequences within these sites. A novel mechanism for the amplification of these SSR sequences appears to be the root cause of these genetic abnormalities. In Kenedy's disease, a $(CAG)_n$ repeat in the coding sequence of an androgen receptor gene increases from the normal copy number of n = 11-31 to n = 40-60. The Huntington's disease gene has a $(CAG)_n$ repeat which increases from a normal copy number of n = 11-34 to n > 50 in cases of the disease (Brown *et al.*, 1996). A similar mechanism occurs with a number of other diseases, such as Fragile-X syndrome, with a $(CGG)_n$ repeat in the coding sequence, and Myotonic dystrophy with a $(CTG)_n$ repeat located at the 3 end of the coding sequence of the myotonic Kinase gene increasing from n = 5-35 in normal genes up to thousands of copies in cases where the disease is manifested. If a similar mechanism works in plants, mapping SSR markers from disease resistance ESTs may increase the probability of tagging resistance genes.

The study of EST-SSR will inevitably yield information on the location and expression of many genes located adjacent to SSR loci. This will be facilitated by sequence comparisons between sequenced SSR regions and sugarcane cDNAs in the huge SUCEST database, and it is probable that the accumulation of sequence information based on SSR analysis of the sugarcane genome will lead to unexpected discoveries about sugarcane genes and genome organization.

Another aspect of the complex sugarcane genome that can be addressed using the information obtained from repetitive DNA is polyploid genome evolution. This occurs in allopolyploid cotton, where repetitive DNA accounts for about half of the genome size differences between the 2 diploid progenitors. Continued application of molecular markers, such as EST-SSRs, to sugarcane genome analysis holds great promise for producing lasting insight into processes by which novel genotypes are generated.

Polyploid plants have increased heterozygosity, an attribute that may be beneficial and has been reported to occur in sugarcane (Ming *et al.*, 1998). Polyploids also harbor higher levels of genetic and genomic diversity than was anticipated (Soltis and Soltis, 2000), resulting in greater biochemical diversity which may also be beneficial to the plant. Both mechanisms, heterozygosity and genetic diversity, are favored by the occurrence of SSRs within DNA coding sequences, which helps to explain the high frequency of SSR within sugarcane ESTs.

EST-SSR technology offers the potential of more cost effective data acquisition than is the case with other technologies. EST-SSR vary in different varieties of sugarcane, and this variability may be used to develop molecular markers for mapping sugarcane genes and traits, SSRs being the part of the sugarcane genome predicted to be most immediately useful to plant breeders and geneticists.

REFERENCES

Brown, S.M., Szewc-McFadden A.K. and **Kresovich, S.** (1996). Development and Application of Simple Sequence Repeat (SSR) Loci for Plant Genome Analysis. In: *Methods of Genome Analysis in Plant* (Jauhar, Prem, ed.). CRC Press, Inc. Boca Raton, Florida, USA, 386 pp.

Coe Jr, E., Cone, K.C. and **Gardiner, J.** (2000). *Maize Genetics Conference Abstracts* 42, 191:W.

Cordeiro, G.M., Maguire, T.L., Edwards, K.J. and **Henry, J.R.** (1999). Optimization of Microsatellite Enrichment Technique in *Saccharum spp*. *Plant Molecular Biology Reporter* 17: 225-229.

Gupta, P.K., Balyan, H.S., Sharma, P.C., and **Ramesh, B.** (1996). Microsatellites in Plants: A new class of molecular markers. *Current Science* 70 (1): 45-54.

Ming, R., Liu, S.C., Lin, Y.R. , da Silva, J.A., Wilson, W., Braga, D., van Deynze, A., Wenslaff, T.F., Wu, K.K., Moore, P.H., Burnquist, W., Sorrells, M.E, Irvine, J. and **Paterson. E.A.H.** (1998). Detail alignment of *Saccharum* and *Sorghum* chromosomes: Comparative organization of closely related diploid and polyploid genomes Genetics 150 (4): 1663-1682.

Scott, K.D., Eggler, P., Seaton, G., Rossetto, M., Ablet, E.M., Lee, L.S. and **Henry, R.J.** (2000). Analysis of SSRs derived from grape ESTs. Theor. Appl. Genet. 100: 723-726.

Soltis, P. and **Soltis. D.E. (**2000). The Role of Genetic and Genomic Attributes in the Success of Polyploids. *PNAS* 97 (13): 7051-7057.

Taramino, G., Tarchini, R., Ferrario, S., Lee, M. and **Pe, M.E.** (1997). Characterization and mapping of simple sequence repeats (SSRs) in Sorghum bicolor. *Theor. Appl. Genet* 95: 66-72.

Wren J.D., Forgacs, E., Fondon 3rd., J.W., Pertsemlidis, A., Cheng, S., Gallardo, T., Williams, R.S., Sorret, R.V., Minna J.D. and **Garner, H.R.** (2001). Repeat Polymorphisms Within Gene Regions - Phenotypic and Evolutionary Implications. *American Journal of Human Genetics* (in press).