J. Hortl. Sci.
Vol. 6(1):37-40, 2011

In silico microsatellite development in arum lily (*Zantedeschia aethiopica)*

V. Radhika, C. Aswath, D.C. Lakshman Reddy, Shweta, A. Bhardwaj
Division of Biotechnology, Indian Institute of Horticultural Research
Hessaraghatta Lake P.O., Bangalore - 560 089, India
E-mail: vr@iihr.ernet.in

ABSTRACT

Microsatellites are an important class of molecular markers having wide application in genetic research. Development of microsatellites using conventional methods is laborious and expensive. Alternatively, *in silico* approach can be followed to detect simple sequence repeats (SSRs) from expressed sequence tags (ESTs) available in public biological databases. The *in silico* developed EST-SSRs have been found to be transferrable across species and genera. A study was undertaken to mine simple sequence repeats (SSRs) from the expressed sequence tags (ESTs) of arum lily, *Zantedeschia aethiopica*, belongs to the family Araceae. A total of 4283 ESTs of *Zantedeschia aethiopica*, downloaded from dbEST of NCBI, were pre-processed and subjected to clustering and assembly. In all, 1968 clusters (800 contigs and 1168 singletons) were obtained, resulting in 54 % reduction in ESTs. In addition, 1936 SSRs were obtained, which included 617 mono, 101 di-, 201 tri-, 80 tetra-, 23 penta- and 898 hexa-nucleotide repeats. The plant has an abundance of 0.70 SSRs/ kb. We designed 1091 primers for these SSRs. A few *in silico* designed SSR primers were tested for polymorphism in *Anthurium*, belonging to the Araceae family, resulting in 40% amplification success.

Key Words: *Anthurium*, Araceae, Expressed Sequence Tag (EST), Microsatellite, Simple Sequence Repeat (SSR)

INTRODUCTION

SSRs, also known as microsatellites, are tandem repeat units that are 1-6 nucleotides in length. They represent an important class of molecular markers for studying the genome structure, evolution and applied aspects. The SSRs have wide applications in plant breeding. Single nucleotide SSRs have been used in population genetics analyses of chloroplast genomes, while di-, tri- and tetra- nucleotide SSRs are used for the construction of linkage maps of nuclear genomes. Development of SSRs using conventional methods by means of construction and screening of genomic libraries, sequencing of clones containing SSRs and testing of primers is time-consuming, expensive and laborious. Alternatively, using bioinformatics approach, SSRs can be mined from the EST sequences available in the genomic databases.

EST projects for several genomes are in progress and are generating a large number of EST sequences for numerous organisms. The ESTs are deposited in the public biological databases and available freely for download. EST sequences represent the coding regions of the genome and hence useful for marker development. SSR mining from ESTs has been explored for monocot crops (Kantety *et al*, 2002; Jayashree *et al*, 2006) and few dicot crops (Kumpatla and Mukhopadhyay, 2005; Scott *et al*, 2000; Jayashree *et al*, 2006). A considerable proportion of SSRs developed from ESTs for a given species have been transferable in related plant species (Cordeiro *et al*, 2001; Eujayl *et al*, 2004; Varshney *et al*, 2005) as well as distant plant species (Decroocq *et al*, 2003; Zhang *et al*, 2005). All these features make SSR development from ESTs attractive.

Anthurium and *Zantedeschia* are large genera of flowering plants from the Araceae family. At the time of study the EST database of NCBI did not contain ESTs of anthurium, but there were a considerable number of ESTs of *Zantedeschia aethiopica*, commonly known as arum lily. Hence, a study was undertaken to mine SSRs from ESTs of *Zantedeschia aethiopica*. These SSRs would be further tested and used in anthurium breeding programme of the institute.

MATERIAL AND METHODS

EST data source

The EST sequences of *Zantedeschia aethiopica* were downloaded from dbEST of GenBank (Boguski *et al*, 1993). 4283 EST sequences were downloaded in fasta format.

EST processing

Pre-processing of the ESTs was carried out in 4 steps - elimination of vector contamination, removal of poly-A tails and ambiguous bits. Freely available tools were used for this purpose. The EST sequences were compared with the Univec vector database using VecScreen (*www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html)* for identification of vector contamination. The TrimEST program of EMBOSS was used for removing the polyA/ polyT ends of the EST sequences.

Clustering and assembly

The pre-processed EST sequences were clustered and assembled using CAP3 software (Huang and Madan, 1999) to obtain clusters (contigs and singletons).

SSR identification and primer development

SSR or microsatellite identification from the contigs and singletons was done using I-Mex microsatellite detection software (Suresh and Hampa Pathalu, 2007). Primers were designed for the SSRs using the Primer3 software accessed through the interface of I-Mex.

RESULTS AND DISCUSSION

Pre-processing of ESTs

The total sequence length of the 4283 EST dataset of *Zantedescia aethiopica* was 2.764 Mb. The EST sequences generated in the laboratory may be contaminated with vector sequence. Many of the ESTs are deposited in the EST databases without removal of the contamination. Hence removal of vector contamination from the EST sequences is necessary to improve the efficacy of subsequent analyses. VecScreen of NCBI was used for detection of vector sequences. The identified vector contaminations as well as suspect sequences were removed from the sequences using perl scripts developed in-house. Poly A/ poly T ends of the EST sequences were removed using TrimEST.

EST clustering and assembly

Redundancy has been observed in the ESTs of *Zantedeschia aethiopica.* 1840 clusters (800 contigs and 1168 singletons) were obtained after clustering and assembly of the ESTs resulting in a 54% reduction of the EST data. Redundancy is an inherent feature with EST datasets generated by random or shotgun sequencing within cDNA libraries. Clustering of ESTs eliminates redundancy in the datasets. CAP3 computes overlaps between sequences and joins reads in decreasing order of overlap scores to form contigs. The contigs are longer in length, which facilitates the design of primers for those EST-SSRs where the SSR is near the end of the EST sequence. The contigs and singletons obtained were used for identification of SSRs.

SSR identification

The SSRs were detected using I-Mex. The minimum number of repeats was fixed as 20 for mono nucleotide, 6 for di and tri nucleotide, 3 for tetra and penta nucleotide and 2 for hexa nucleotide repeats. Further the analysis of occurrence and frequency of SSRs was investigated to find repeat types, number of repeats, and frequency. The definition of SSR varies by size and type of repeat and some authors do not consider monomer repeats as SSRs. Few authors consider only those SSRs whose repeat motif is larger than 20 bp (Varshney *et al,* 2002). Kumpatla and Mukhopadhyay (2005) observed that comparisons of SSR size and type of repeat are difficult to discuss. They targeted four classes of repeats (viz. mono, di, tri and tetra) with default settings for repeats as 15 for mono nucleotides and five for di, tri, or tetra-nucleotides.

1936 SSRs were obtained in *Zantedeschia aethiopica* satisfying the criteria of minimum number of repeats as defined above. This plant has abundance (number of SSRs per kb of sequence analyzed) of 0.70 SSRs/kb which was higher than observed in citrus viz. 0.2 SSRs/Kb of ESTs sequences (Chen *et al,* 2005).

In our study, hexa–nucleotide repeats were the most frequent (46.38%), followed by mono-repeats (31.86%) and then by tri-repeats (11.2%) (Table 1). In citrus and jatropha (Wen *et al,* 2010) the tri-nucleotide motifs were the most abundant while di nucleotide repeats constituted the highest number of repeats in iris (Tang *et al,* 2009).

In all types of SSRs, most common and longest SSRs were found. Most common di nucleotide repeats contained AG motifs. According to past studies, GA/ CT have been found to be the most abundant motifs in barley, maize, sorghum, wheat (Kantety *et al,* 2002), iris (Tang *et al,* 2009) and coffee (Hendre *et al,* 2008).

In the present study, TCT, CTC, GAA were the most

Table 1. Frequency and density of EST-SSRs

Type of repeat	No. of SSRs	Frequency (SSRs/ Kb)	Density (bp/ Mb)
Mono-nucleotide	617	0.2231	9461.65
Di-nucleotide	101	0.0365	329.59
Tri-nucleotide	217	0.078	357.09
Tetra-nucleotide	80	0.0289	93.342
Penta-nucleotide	23	0.0083	26.41
Hexa-nucleotide	898	0.3248	663.53

common tri-nucleotide repeats. Kantety *et al* (2002) found GGC/ CCG the most abundant tri-nucleotide repeat motif in rice, barley, maize and sorghum and AAC/ TTG the most common tri-nucleotide repeats in wheat. CAT/ ATG and TTC/ GAA were the most abundant tri-nucleotide repeats in coffee (Hendre *et al,* 2008) and AAG/ CTT and AGG/ CCT were the most abundant in *Iris (*Tang *et al,* 2009*)*.

The tri-nucleotide repeats act as amino acid codons. Most commonly found were Leucine, Proline, Glycine, Arginine, Glutamic acid, Glycine and Alanine. Most of the codons were repeated 2-3 times. Amino acid repeats for Leucine and Alanine were most frequent in the ESTs (Table 3).

Most common tetra nucleotide repeat in *Zanthedeschia aethiopica* was TCCC. The lengths of tri-nucleotide and tetra-nucleotide repeat in this plant ranged from 18-31 and 12-28 repeats respectively.

The SSR loci were categorized into two groups based on the length of their SSR tract size, class I having SSRs length >= 20 and class II containing perfect SSRs > 12 but < 20. Of the total number of SSRs identified, 676 belonged to class I and 142 to class II (Figure 1). Mono-nucleotide repeats formed the largest portion of class I and tri-nucleotide repeats formed the largest portion of class II repeats. Class I and class II microsatellites were found to be most frequent in the gene rich regions in rice with a higher frequency of class II than class I repeats (Temnykh *et al,* 2001).

Primer development

To convert *in silico* identified EST-SSRs to molecular markers, primer pairs were designed for EST-SSRs. Primers were designed for the SSRs using Primer3 software. Primer pairs could be designed for 70% (1091) of the EST-SSRs.

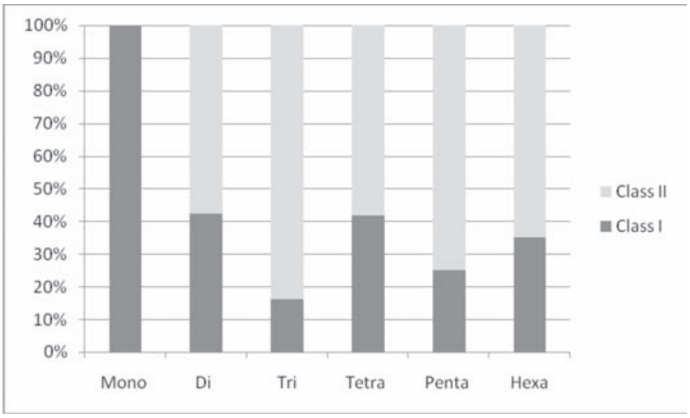

Fig 1. Distribution of class I and class II repeats in arum lily, *Zantedeschia aethiopica*

Table 2. Most common and longest SSR motifs

Nucleotide Repeat type	Most common repeat/s	Frequency	Longest repeat
Di	AG	15	TC (50)
Tri	TCT, GAA, CTC	24	CTT (30)
Tetra	TCCC	4	AAAT (28)
Penta	-	-	GTTGC (25) CCCTC (25)
Hexa	CTCGGA	10	CATCAA (36) AAAAAG (36)

Table 3. Frequency of tri-nucleotide repeats as EST-SSRs

Amino acid	Codons	Frequency
Alanine	CGA, GCG, GCC, GCT	30
Arginine	AGA, AGG, CGC, CGG, CGT	21
Asparagine	AAC	1
Cysteine	TGC, TGT	8
Glutamic Acid	GAA, GAG	13
Glutamine	CAA,CAG	9
Glycine	GGC, GGT, GGA	29
Histidine	CAT, CAC	7
Isoleucine	ATA, ATT	2
Lysine	AAG	6
Leucine	CTC, CTG, TTA, CTT, CTA, TTA	30
Aspartic Acid	GAC, GAT	3
Methionine	ATG	1
Serine	AGC, TCA, TCC, TCT	22
Phenyl alanine	TTC	3
Proline	CCA, CCT, CCG	13
Stop codon	TGA	5
Theronine	ACA, ACC, ACT	4
Tryphtophan	TGG	5
Valine	GTT, GTG, GTA	5

Few EST-SSRs of *Zantedeschia aethiopica* were tested for polymorphism in *Anthurium* and 40% amplification success was obtained. In coffee, 61 (63.5%) primer pairs were experimentally validated and used to investigate the genetic diversity among the 34 accessions of different *Camellia* spp. (Hendre *et al,* 2008). The level of polymorphism in cultivars detected by EST-SSR markers in sugarcane was low (PIC=0.23) while a subset of these markers showed a high level of polymorphism in related genera viz. erianthus and sorghum species (Cordeiro *et al,* 2001). The primers pairs designed for EST-SSRs of *Medicago truncatula* showed high level of polymorphism (70%) in alfalfa and other annual medics and are valuable genetic markers for the *Medicago* (Eujayl *et al*, 2004). A subset of 165 EST-SSR markers from a total of 185 assigned to genetic map of barley showed transferability in wheat (78.2%), rye (75.2%) and in rice (42.4%) (Varshney *et al,* 2005). The transferability of EST-SSR markers from apricot and grapevine to other related and unrelated species was examined (Decroocq *et al,* 2003). Overall grape primers amplified products in most of Vitaceae accessions while apricot primers amplified polymorphic alleles only in closely related species of Rosaceae. Transferability of EST-SSRs of *Triticum aestivum* was studied in eight related species and it ranged from 76.7 % for *Aegilops tauschii* to 90.4% for *T. durum* and was lower for distant relatives such as barley (50.4%) and rice (28.3%) (Zhang *et al*, 2005).

ACKNOWLEDGEMENT

The authors would like to thank Director, Indian Institute of Horticultural Research, Bangalore for encouragement. This research was partly funded by Council of Scientific and Industrial Research (CSIR), New Delhi, India.

REFERENCES

Boguski, M.S., Lowe, T.M. and Tolstoshev,C.M. 1993. dbEST—database for “expressed sequence tags”. *Nat Genet.*, **4**:332-333

Chen, C., Zhou, P., Choi, Y.A., Huang, S. and Gmitter, F.G. 2005. Mining and characterizing microsatellites from citrus ESTs. *Theor. Appl. Genet*., **112**: 1248- 1257

Cordeiro, G. M., Casu, R., McIntyre, C.L., Manners, J.M. and Henry, R.J. 2001. Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to *Erianthus* and sorghum. *Pl. Sci.,* **160:**1115-1123

Decroocq, V., Fave, M.G., Hagen, L., Bordenave, L. and Decroocq, S. 2003. Development and transferability of apricot and grape EST microsatellite markers across taxa. *Theor. Appl. Genet.,* **106:**912-922

Eujayl, I., Sledge, M., Wang, L., May, G.D., Chekhovskiy, K., Zwonitzer, J.C. and Mian, M.A.R. 2004. *Medicago truncatula* EST-SSRs reveal cross-species genetic markers for *Medicago* spp. *Theor. Appl. Genet.*, **108:**414-422

Hendre, P.S., Phanindranath, R., Annapurna, V., Lalremruata, A. and Aggarwal, R.K. 2008. Development of new genomic microsatellite markers from robusta coffee (*Coffea canephora* Pierre ex a. Froehner) showing broad cross- species transferability and utility in genetic studies. *BMC Pl Biol,* **8:**51

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

Jayashree, B., Punna, R., Prasad, P., Bantte , K., Has, C.T., Chandra, S., Hoisington, D.A. and Varshney, R.K. 2006. A database of simple sequence repeats from cereal and legume expressed sequence tags mined *in silico*: survey and evaluation. *In Silico Biol*, **6**: 607-20

Kantety, R.V., Rota, M. L., Matthews , D.E. and Sorrells, M.E. 2002. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Pl. Mol. Biol.,* **48**:501-510

Kumpatla, S.P. and Mukhopadhyay, S. 2005. Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. *Genome*, **48**:985-98

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

Suresh, B.M. and Hampapathalu, A.N. 2007. IMEx: Imperfect Microsatellite Extractor. *Bioinformatics* **23**:1181-1187

Tang, S., Okashah, R.A., Pratt, M.M.C., Pratt, L.H., Johnson, V.E., Taylor, C., Arnold, M.L. and Knapp, S.J. 2009. EST and EST-SSR marker resources for *Iris. BMC Pl. Biol.*, **9**:72

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

Varshney, R.K., 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. Lett.,* **7**:537–546

Varshney, R.K., Sigmund, R., Borner, A., Korzun, V., Stein, N., Sorrells, M.E., Langridge, P. and Graner, A. 2005. Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice *Pl. Sci.,* **168:**195-202

Wen, M., Wang ,H., Xia, Z., Zou, M., Lu, C. and Wang, W. 2010. Development of EST-SSR and genomic-SSR markers to assess genetic diversity in *Jatropha curcas* **L.** *BMC Res. Notes.,* **3**:42

Zhang, L.Y., Bernard, M., Leroy, P., Feuillet, C. and Sourdille, P. 2005. High transferability of bread wheat EST-derived SSRs to other cereals. *Theor. Appl. Genet.,* **111**:677-687

(MS Received 20 October 2010, Revised 25 February 2011)