Abrahamia
Vol. 2
51-56
2016
ISSN: 2395-4108

Development and validation of EST-SSR and identification of EST-SNP markers for snake gourd (*Trichosanthes cucumerina var. cucumerina L.*)

Silpa Sivan[1], Shinsy M. Y.[2] & Sabu K. K.[3]*

[1]Department of Computational Biology and Bioinformatics, University of Kerala, Thiruvananthapuram 695581.
[2]Department of Bioinformatics, Union Christian College, Aluva 683102.
[3]Biotechnology and Bioinformatics Division, Jawaharlal Nehru Tropical Botanic Garden and Research Institute, Thiruvananthapuram 695562, India.
*Corresponding author Email: sabu@jntbgri.res.in

Abstract

Snake gourd is a natural antibiotic and the plant is commonly used in folk medicines throughout India. The present study was carried out using transcriptomic sequence of *Trichosanthes kirilowii* Maxim. for the development and validation of EST-SSR markers and EST sequences of *T. dioica* Roxb. for identification of EST-SNPs in *T. cucumerina* var. c*ucumerina* L. A total of 382 primers were developed and out of which 58 were found to be suitable for genetic analysis. For experimental validation, 20 primers were randomly selected and amplified in *T. cucumerina* var. c*ucumerina* and two related genera, *viz. Coccinia grandis* and *Momordica charantia*. It was found that 90% of the primers generated PCR products in *T. cucumerina* var. c*ucumerina* and considerable cross generic amplification in *M. charantia* and *C. grandis.* Further, identification of EST-SNPs in *T. cucumerina* var. c*ucumerina* was also carried out. The SNPs detected were converted to CAPS markers using the tool SNP2CAPS. The newly developed SSR and SNP markers offer novel additions to the genomic resources of this important medicinal herb for population studies, phylogenetics analyses and estimation of genetic diversity.

Keywords: EST-SSR, EST-SNP, SNP2CAPS, *Trichosanthes cucumerina* var. c*ucumerina*

Introduction

Trichosanthes cucumerina var. *cucumerina* L. (Cucurbitaceae) is a climber and is cultivated for its long fruit. It possesses high nutritional value and medical importance due to the abundance of flavonoids, carotenoids, phenolic acids, etc. The plant has pharmacological and therapeutical activities such as anti-diabetic, hepatoprotective, cytotoxic, anti-inflammatory, etc and it is used in Ayurveda and Siddha medical practices (Warrior *et al*., 1993). However, there is no suitable variety of *T. cucumerina* var. c*ucumerina* developed for cultivation and hence a large number of locally available plants are cultivated in every part of the country. In order to carry out varietal improvement, germplasm should be screened for genetic variability along with desirable characteristics such as disease resistance, stress tolerance, yield, earliness, etc (Choudhury *et al*., 1975).

Since DNA markers are not affected by external environmental conditions, it can be used as marker of choice for genetic variation studies among plant populations (Jubera *et al*., 2009). Data provided by molecular markers can be used for selection during backcross breeding programs (Varshney *et al*., 2005). Marker Assisted Selection (MAS) has proven that it is the best method for improvement of many crop plants (Raji *et al*., 2009). The search of molecular markers such as SSRs (simple sequence repeats) in ESTs (expressed sequence tags representing genes or coding regions) are widely used because of the availability of higher number of SSRs in single or low-copy (Morgante *et al*., 2002; Anjali *et al*., 2016) and which can be used for designing locus-specific primers (Riju *et al*., 2007). The advantages of developing EST-SSR markers are low cost, technical simplicity, and power to capture the functional diversity in natural populations (Varshney *et al*., 2005; Chakravarthi

and Naravaneni, 2006). Other popular markers like INDELs (insertions and deletions) and SNPs (single nucleotide polymorphisms) are also increasingly used for genetic diversity analysis in crop plants (Gupta and Varshney, 2000).

Understanding the genes that are responsible for a particular trait will be useful for crop improvement. Several bioinformatic tools are available to select plants with desirable quality, or with the ability to survive on extreme environmental conditions. Analysis of transcriptome, the complete set of gene transcripts in a cell, is helpful for studying the genome and revealing the expression of molecular constituents of cells and tissues, particularly for medicinal plants (Zong *et al*., 2009; Lakshmi Priya and Sabu, 2015). Large number of EST-SSRs (Bouck and Vision, 2007) and SNPs (Cloonan *et al.*, 2008) can be generated through transcriptome sequencing using high-throughput methods followed by bioinformatics analysis (Zhong *et al.*, 2009; Nadiya *et al*., 2017).

Progress as well as interest in plant genomics has greatly increased with the array of genome sequencing projects that had been started after the publication of genome of *Arabidopsis thaliana* in late 2000. However, a combination of bioinformatic tools of genomic annotation and comparative genomics have to be used to get an understanding of any genome (Gary and Chunguang, 2008; Lakshmi Priya and Sabu, 2015; Thomas and Sabu, 2016). The present investigation was aimed to develop and validate EST-SSR and identify EST-SNP markers for analyzing the genetic variability in snake gourd (*Trichosanthes cucumerina* var. *cucumerina* L.) through bioinformatics analysis of the transcriptome of *T. kirilowii* and ESTs of *T. dioica* respectively.

Materials and Methods

Transcriptome sequence reads of *Trichosanthes kirilowii* Maxim. were accessed from NCBI's SRA database (http://www.ncbi.nlm.nih.gov/sra/) under accession number SRX 1078569. Quality checking of the raw reads were carried out using FASTQC toolkit (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). *De novo* assembly of the individual reads to full length transcriptome was accomplished through DDBJ (http://www.ddbj.nig.ac.jp/). Identification of the ESTs in the assembled transcriptome was performed using a perl script, EST_trimmer.pl (pgrc.ipk-gatersleben.de/misa/download/est_trimmer.pl). Identification of the SSRs in the ESTs was carried out using Micro Satellite Identification Tool (MISA v1.0). Primers to amplify the SSR regions were designed using WebSat (wsmartins.net/websat/). The quality of the primers were assessed using Net Primer (www.premierbiosoft.com/netprimer/), where the primers were analyzed for all secondary structures including hairpins, self-dimers, and cross-dimers in primer pairs. A rating of hundred percent indicates the absence of self and cross dimerization of the primers. Twenty primers that showed 100% rating were synthesized at Sigma-Aldrich, Bengaluru and used for wet lab validation.

Genomic DNA was isolated from young leaves of *Trichosanthes cucumerina* var. *cucumerina, Coccinia grandis,* and *Memordica charantia* using DNA Purification Kit (Origin Labs, Kerala). Isolated DNA samples were loaded onto 0.8% agarose gel, electrophoresed at 70V for 45min and examined under UV light in a gel documentation system (UVP, UK) to find out the integrity. The quality of the isolated genomic DNA was also assessed using BioPhotometer (Eppendorf India Ltd). Polymerase chain reactions were carried out in Veriti® Thermal Cycler (Applied Biosystems, USA). The final volume of 25 µl PCR reaction mixture comprised 50ng of genomic DNA, 200 µM of each four dNTPs, 15 picomole of both forward and reverse primers, 1x Taq buffer and 1 unit Taq polymerase enzyme. The PCR conditions were: 94°C for 2 min, followed by 35 cycles of 94°C for 30 sec, specific annealing temperature for 1 min, 72°C for 2 min, and 72°C for 7 min for the final extension. The amplified products were separated using agarose (3%) in horizontal electrophoresis unit at 70V and visualized by ethidium bromide under UV using the gel documentation system. The bands obtained were scored by following expected product size as reference.

For identification of EST-SNPs, EST sequences of *T. dioica* Roxb. were obtained from NCBI-EST database under accession numbers EX015356.1, EX015355.1, EX015354.1 and EX015353.1. These EST sequences were downloaded in FASTA format and subjected to multiple sequence alignment using Clustal W (www.genome.jp/tools/clustalw). Detection of the SNPs and conversion of the same to CAPS (cleaved amplified polymorphic sequence) markers were achieved through SNP2CAPS tool (http://pgrc.ipk-gatersleben.de/snp2caps/). Two input files that contain data on the sequence alignments and the restriction enzymes are required for SNP2CAPS. The first input file was a modified FASTA formatted file that stored multiple alignments of sequences of *T. dioica*. The second input file contained data on the restriction enzymes downloaded from

the restriction enzyme database REBASE (http://rebase.neb.com).

Results and Discussion

Several medicinal plants that are used in the traditional medical practices are increasingly being validated to understand their chemical constituents or systematically analyze their efficacy (Bansode *et al.,* 2016; Pluhár *et al.,* 2016). Improvement of the yield and quality of these natural resources through conventional breeding is still a challenge. However, with the recent advances in plant genomics, considerable progress has also been made in the development of functional genomics resources (EST/SNP databases and micro-arrays) in several medicinal plant species, which offer new opportunities for improvement of genotypes using perfect markers or genetic transformation (Kumar and Gupta, 2008).

The raw transcriptomic sequence of *Trichosanthes kirilowii* was used to develop EST-SSR markers as the transcriptome of *T. cucumerina* var. c*ucumerina* was not available. FASTQC analysis indicated that total number of sequence was around 20 million, sequence length was 100 and the GC content was 45%. The base sequence quality score and sequence quality were normal for further analyses. No overrepresented sequences and adapters were identified and hence the transcriptomic sequence was of good quality for further analyses. The sequences with a total size of 5.5 Gb was *de novo* assembled using Trinity program in the DDBJ server resulting in the assembled transcriptome of 169 Mb. After assembling, the assembled sequence was converted to expressed sequence tag contig file. After trimming, the sequence size was reduced to 52.2Mb. A total of 23,950 SSRs were identified using the MISA tool in 20,069 sequences. Number of sequences containing more than one SSR was 3,223. Several types of SSR repeats were identified in the EST contig where the percentage distribution of mono nucleotides was much higher in comparison to di, tri, tetra, penta, and hexa nucloetides (Fig 1).

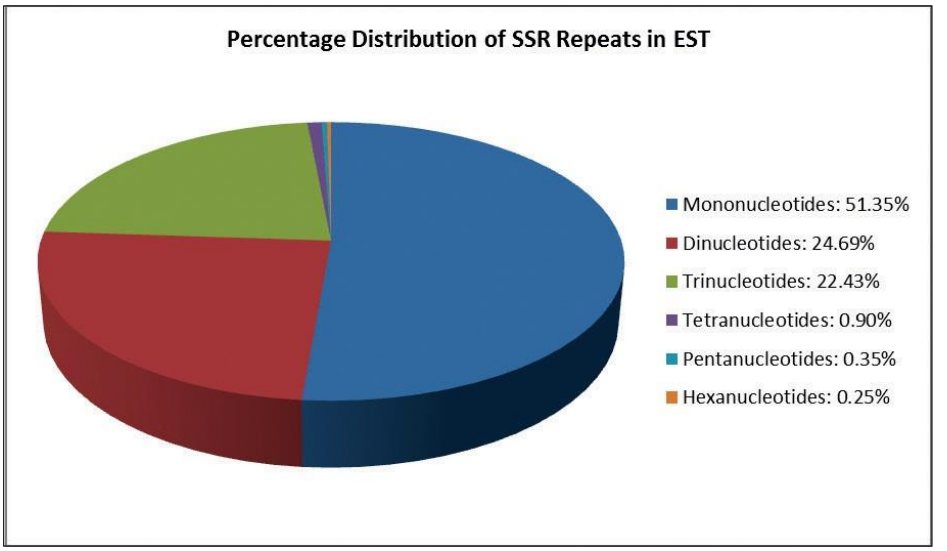

Fig. 1 Percentage distribution of SSR repeats in the EST contig

As reported earlier (Jain *et al.*, 2014), mononucleotide repeats were not considered for marker development since they do not have any importance in studies involving molecular markers and it was also found that they cause difficulties in accurate sizing of polymorphisms. Hence, to amplify the SSR repeats excluding the mono nucleotides, 382 primers were designed using Websat and out of which 58 primers were found as having 100% efficiency. From these 20 primers (Table 1) were selected for experimental validation in *Trichosanthes cucumerina* var. *cucumerina* and two related genera such as *Coccinia grandis,* and *Memordica charantia*. The gradient temperature set for *T. cucumerina* var. *cucumerina* was used in the other two species. About 80% annealing temperature which was set for *T. cucumerina* var. *cucumerina* worked for *M. charantia* and 60% for *C. grandis*. 2 SSR primers did not work on any of the annealing temperatures within the range of 52-65˚C. In the case of some of the SSR primers the expected product size was different from that of estimated size.

Table 1.
Sequence information, expected product size and annealing temperature of the selected primers

No.	Forward primer	Reverse primer	Product size (bp)	Annealing Temperature (˚C)
1	CTCCCACTCTCTCTCTCT CT GTCT	ACTCCCATTCACAAGTTTACC G	275	55.1
2	CTGAGAATGTGGGAGCAAA AC	GATACCAAGCATAGCAGAAC CC	386	56.5
3	GAGAAGAGAGCAAGAC CGAGAC	AGGCACTTGACAGAGACGACTT	386	-
4	GTTGGAGTTCCTGTGGTCTTT C	TCTCTCTCTCCCTTACCGTTC A	205	-
5	TCACAACACCTCCTCTCTCTC TC	TCACGGTCAATCACTTTCAAC T	262	64.2
6	TCTATCTGGTCTACTGGGGAG C	CATACACACACAAAACCACCC T	357	55.1
7	CCAACACTCAAAATCACC AA GA	ACTCCTATCAAATGGGGTTCC T	155	64.2
8	CACACACACACACACA CACTCTC	GTCGTCATCTCCTTCTTGCTC T	175	52.9
9	CATCTTCGTCCTCAACCTCA TC	GGTCGTCATTATCGTCGTCTT T	135	55.1
10	CCAGATTGGAAATAGAGGA G GA	AAGCGTGGAGAAGAGACAAAA G	302	56.5
11	AGAAAGGGAAATGGAAGA G GAG	AGATGATGGTGATGAGATGA CG	237	53.3
12	CGATTGAGGAGAA G GAGAAGAA	AACAAACAACACCAACATCGT C	115	53.5
13	AGGAGGATTTGAGTTGGGAT GT	CCTCTTTCACTGTCTCACCGA T	114	55.8
14	CTGATGGATTTGGTTCTGGTT T	TCCTGATTTGGGTTCTTCTTG T	400	54.4
15	CTGTTTCTTTTCTGTTTTGGG G	ACCACCTTTGTTGTTGTTGTT G	355	53.5
16	TCAAAGAGAGAGAGC GAAATCC	TTCGGAGAAGAAGAAGACAAG G	163	53.5
17	AAGAAGAAGAAGGGGAAAA T GG	CGTGGAGTCTGAAACCGAAT	223	56.9
18	GACGAGACAACGCAGATAGG A	CAACCTAAAATCTCAAATCCC G	109	53.5
19	GATTCCAGATTCATCGGTTC AC	ATCAGGGAGGGAGGAAGGA	124	57.7
20	TCTGCTTGGAGTTGGTATTC G	CTGAGAAGATAAGATTGATTA C GGG	194	52.5

A single nucleotide polymorphism is a variation in a single nucleotide (A, T, G or C) that occurs at a specific position in a genome. For the analysis, each of the aligned sequences was screened for the corresponding restriction enzyme recognition sites using the standard regular expression syntax. Only those enzymes that display a restriction site polymorphism among all of the aligned sequences were considered for the marker development. CAPS candidates falls into two categories- alternative restriction patterns or may occur because of either SNPs or InDels. In the case of SNPs, the algorithm decides to assign the recognition site to one of the four classes. For computational restriction analysis the GCG format data file from the REBASE database (version 608, July 29, 2016) was used. It contained information on 934 type II restriction enzymes. For our study, we selected (a) total of 200 enzymes that are non-isoschizomers and commercially available and (b) a subset of 10 enzymes (Table 2). The SNP markers identified can be experimentally validated with the set of common restriction enzymes. Upon restriction digestion of corresponding PCR fragments, these EST-restriction enzyme pairs can be tested that reveals the computationally predicted restriction pattern.

Table 2.
Conversion of SNPs to putative CAPS markers

CAPS candidates	200 Enzymes from RE Database	10 randomly selected enzymes from RE Database
Predicted CAPS Candidates	252	80
Predicted CAPS Candidate containing N	5	5
False positive CAPS Candidates	8	8
No restriction site polymorphism	28	130

Conclusion

Snake gourd is a relatively well-known veggie utilized in numerous standard recipes. In Ayurveda, the snake gourd is a cure for intestinal infestations. In this work, attempts were made to develop and validate EST-SSR markers and to identify EST-SNPs in *T. cucumerina* var. *cucumerina* and its related genera using transcriptomic sequence of *T. kirilowii* and *T. dioica* respectively. The EST-SSR primers showed 90% amplification in *T. cucumerina* var. *cucumerina* and also exhibited significant amount of cross generic amplification. Similarly, the EST-SNPs were also successfully identified in the study. The newly developed markers can be used for population studies, phylogenetics analyses and genetic diversity of *T. cucumerina* var. *cucumerina* and related genera.

Acknowledgements

Authors thank Director, JNTBGRI for the facilities extended to carry out this research programme. SS and SMY acknowledge their gratitude to members of the Plant Molecular Biology Laboratory of Biotechnology and Bioinformatics Division, JNTBGRI for all academic and technical help.

Authors are deeply indebted to Jinu Thomas for assistance offered to develop the bioinformatics pipeline required for the marker development.

References

Anjali, N., Sowmya S. Dharan, Nadiya, F., Sabu, K. K. (2015) Development of EST-SSR markers to assess genetic diversity in *Elettaria cardamomum* Maton. International *Journal of Applied Sciences and Biotechnology* 3:188-192.

Bansode, T., Gupta, A., Salalkar, B. (2016) *In silico* and *in vitro* assessment on antidiabetic efficacy of secondary metabolites from *Syzygium cumini* (L.) Skeels. *Plant Science Today* 3:360-367.

Bouck, A., Vision, T. (2007) The molecular ecologist's guide to expressed sequence tags. *Molecular Ecology* 16:907 – 924

Chakravarthi, B. K., Naravaneni, R. (2006) SSR Marker Based DNA Fingerprinting and Diversity Study in Rice (*Oryza sativa*. L.). *African Journal of Biotechnology* 5:684-688

Choudhary, B. (1957) *Vegetables National Book Trust*, New Delhi, India. pp. 52-154

Choudhury, B. D., Bhat, P. N., Singh, V. P. (1975) Genetic diversity in cluster beans. *Indian Journal of Agricultural Sciences* 45:530-535

Cloonan, N., Forrest, A. R., Kolle, G., Gardiner, B. B., Faulkner, G. J., Brown, M. K., Taylor, D.

F., Steptoe, A. L., Wani, S., Bethel, G., Robertson, A. J., Perkins, A. C., Bruce, S. J.,, Lee, C. C., Ranade, S. S., Peckham, H. E., Manning, J. M., McKernan, K. J., Grimmond, S. M. (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods* 5:613–619.

Gary, R., Skuse, Chunguang Du (2008) Bioinformatics Tools for Plant Genomics. *International Journal of Plant Genomics*. Hindawi Publishing Corporation. Article ID 910474, 2.

Gupta, P. K., Varshney, R. K. (2000) The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. *Euphytica* 113:163–185.

Gupta S., Prasad, M. (2009) Development and Characterization of Genic SSR Markers in *Medicago truncatula* and Their Transferability in Leguminous and Non-Leguminous Species. *Genome* 52: 761-771.

Jain, N., Patil, G., Bhargava, P., Nadgauda, R. (2014) *In Silico* Mining of EST-SSRs in *Jatropha curcas* L. towards Assessing Genetic Polymorphism and Marker Development for Selection of High Oil Yielding Clones. *Am J Plant Sci.* 5:1521–41.

Jubera, M. A., Janagoudar, B. S., Biradar, D. P., Ravikumar, R. L., Koti, R. V., Patil, S. J. (2009) Genetic Diversity Analysis of Elite *Jatropha curcas* (L.) Genotypes Using Randomly Amplified Polymorphic DNA Markers. *Journal of Agricultural Sciences* 22: 293-295.

Kumar, J., Gupta, P.K. (2008) Molecular approaches for improvement of medicinal and aromatic plants. *Plant Biotechnology Reports* 2:93.

Lakshmi Priya, P. M., Sabu, K. K. (2015). *In Silico* Analysis of Transcriptomes of *Catharanthus roseus* and *Rauvolfia serpentina*, two potent medicinal plants using a pipeline developed from publicly available tools. *American Journal of Bioinformatics Research* 5: 21-25.

Morgante, M., Hanafey, M., Powell, W. (2002) Microsatellites Are Preferentially Associated with Nonrepetitive DNA in Plant Genomes. *Nature Genetics* 30: 194-200.

Nadiya, F., Anjali, N., Thomas, J., Gangaprasad, A., Sabu, K. K. (2017). Transcriptome profiling of *Elettaria cardamomum* (L.) Maton (small cardamom). *Genomics Data* 11: 102–103.

Pluhár, Z., Szabó, D., Sárosi, S. (2016) Effects of different factors influencing the essential oil properties of *Thymus vulgaris* L. *Plant Science Today* 3:312-326.

Raji, A. A. J., Anderson, J. V., Kolade, O. A., Ugwu, C. D., Dixon, A. G. O., Ingelbrecht, I. L. (2009) Gene-Based Microsatellites for Cassava (*Manihot esculenta* Crantz): Prevalence, Polymorphisms, and Cross-Taxa Utility. *BMC Plant Biology* 9: 1471-1429.

Riju, A., Chandrasekar, A., Arunachalam, V. (2007) Mining for Single Nucleotide Polymorphisms and Insertions /Deletions in Expressed Sequence Tag Libraries of Oil Palm. *Bioinformation* 2:128-131.

Thomas, J. Sabu, K. K. (2016). Development of a New Pipeline for Identification and Characterization of MicroRNAs from Plants. *International Journal for Computational Biology* 5:21-27.

Varshney, R. K., Graner, A. Sorrells, M. E. (2005) Genic Microsatellite Markers in Plants: Features and Applications. *Trends in Biotechnology* 23:48-55.

Warrier, P. K., Nambiar, V. P. K., Ramankutty, C. (1993) *Indian medicinal plants compendium of 500 species.* Orient Longman Pvt. ltd. Chennai, pp. 320-322.

Received: 2.8.2016

Revised and Accepted: 12.9.2016