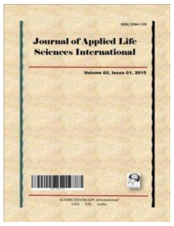
ISSN: 2394-1103

Journal of Applied Life Sciences International

Volume 02, Issue 01, 2015

SCIENCEDOMAIN *international*
USA UK India

Journal of Applied Life Sciences International
2(4): 189-196, 2015; Article no.JALSI.2015.020
ISSN: 2394-1103

SCIENCEDOMAIN *international*
www.sciencedomain.org

Analysis of Frequency and Functional Annotation of SSRs in ESTs of Cole Crops

Rajinder Kaur[1] and Era Vaidya[1*]

[1]Department of Biotechnology, Dr. Y. S. Parmar University of Horticulture and Forestry, Nauni,
Solan-173230, Himachal Pradesh, India.

Authors' contributions

This work was carried out in collaboration between both authors. Author RK designed the study, wrote the protocol and guided the study. Author EV carried out the literature searches, experiments and analyses of the study and wrote the first draft of the manuscript. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/JALSI/2015/11873
Editor(s):
(1) Hayet Hammami, Fungal and Parasitic Molecular Biology Laboratory, Sfax University, Tunisia.
(2) Martin Koller, University of Graz, Research Management and Service, c/o Institute of Chemistry, Austria.
Reviewers:
(1) Anonymous, King Abdulaziz University, KSA.
(2) Feng-De Wang, Shandong Key Laboratory of Greenhouse Vegetable Biology, Shandong Branch of National Vegetable Improvement Center, Institute of Vegetables, Shandong Academy of Agricultural Sciences, Jinan, 250100, China.
(3) Anonymous, Indian Institute of Advanced Research, India.
(4) Anonymous, Brazil.
Complete Peer review History: http://www.sciencedomain.org/review-history.php?iid=880&id=40&aid=9190

Short Research Article

Received 8th June 2014
Accepted 1st April 2015
Published 8th May 2015

ABSTRACT

Aims: *In silico* analysis of publicly available set of expressed sequence tagged sites of *Brassica oleracea* (*B. oleracea* var. *botrytis*, *B. oleracea* var. *capitata*, *B. oleracea* var. *italica* and *B. oleracea* var. *gemmifera*) was carried out so as to remove redundancy. The assembled sequences were then used to detect SSR motifs present among the EST sequences.

Study Design: Detection and annotation of SSRs in EST sequences.

Place and Duration of Study: Department of Biotechnology, Dr Y. S. Parmar University of Horticulture and Forestry, Nauni, Solan-173230, Himachal Pradesh, India, between July 2009 and August 2010.

Methodology: EST sequences were assembled into a non-redundant set using EGassembler program. The assembled sequences were analysed for SSR motifs. The detected EST-SSRs were then subjected to sequence identification and annotation.

Results: After sequence assembly, 34.12% reduction of data redundancy was obtained. 107 SSRs

**Corresponding author: Email: vaidya.era@gmail.com;*

were identified using SSRIT and they were analysed for their distribution in terms of frequency, density, composition and distribution. Maximum frequency was exhibited by dinucleotide SSR motifs i.e. 68.22%. The detected EST-SSRs were annotated for their putative function using BLASTx program of NCBI. 82% SSR-ESTs were found to show a significant match with protein databases. The 100 SSR-ESTs were also analysed for significant functional domains using PROSITE and 83% sequences were found to code for functional domains. Functional domain markers (FDM) predict the functional property of markers having predicted protein domains Primer sequences were designed for some of these EST – SSRs by making use of PRIMER3 software and they were validated for their usefulness on a set of 20 genotypes of *Brassica oleracea* var. *botrytis.* The primers generated high level of genetic polymorphism which could be used for genetic characterization of the genotypes used.

Conclusion: Thus it can be concluded that these detected SSRs can be used to design primers that have functional role and will also facilitate studies on genetic diversity, variability, genome analysis and evolutionary relationships among these cole crops of family *Brassicaceae*.

Keywords: EST; SSRs; cole crops; data mining.

1. INTRODUCTION

Microsatellites or simple sequence repeats (SSRs) are short repeats of 1–6 base pairs in length [1,2], dispersed throughout the genomes of most eukaryotic organisms. SSR primer pairs (forward and reverse) are based on conserved flanking regions and the SSR polymorphisms are based on variation in the number of specific repeat units. SSR markers are multi-allelic, co-dominant, reproducible and hypervariable with wide genome coverage. Microsatellites have been characterized in different crop species including cereals, legumes, fruits and vegetable crops [3-6].

Generally SSRs are isolated and characterized from genomic DNA, but it is a tedious job. However, many reports have demonstrated that a large number of SSRs are located in transcribed regions of genomes, including protein-coding genes and expressed sequence tags (ESTs) [7]. Computer based applications have been used to extract out SSRs from the increasing EST sequences in public databases [8-10].

Advances in bioinformatics tools have aided significantly in generating a large number of ESTs in various crop plants. Recently large datasets of ESTs have been developed that have aided significantly in whole genome sequencing and further annotation of the specific sequences [1,4,11].

The present study was aimed at obtaining ESTs from *Brassica oleracea* var. *botrytis*, *Brassica oleracea* var. *capitata*, *Brassica oleracea* var. *italica* and *Brassica oleracea* var*. gemmifera*, available in public database. These were then further processed as a collective unit to analyze SSRs present in the ESTs and to characterize and annotate the newly detected SSRs.

2. MATERIALS AND METHODS

2.1 Retrieval of EST Sequences

All the EST sequences belonging to four different botanical varieties of *Brassica oleracea* (*B. oleracea* var. *botrytis, B. oleracea* var. *capitata, B. oleracea* var. *italica* and *B. oleracea* var. *gemmifera*), were acquired in their FASTA format from the EST database on the NCBI website. A total of 2564 sequences were retrieved, which originated different plant tissues such as leaves, stem and root. A single text file containing all the 2564 EST sequences was compiled.

2.2 EST Sequence Assembly

The 2564 redundant ESTs were used to produce a non-redundant dataset for clustering and assembly analysis. EG assembler web server [12] was used for automatically removing sequences having low quality or complexity. Small RNA pseudo genes, LINEs, SINEs, LTR elements, vector sequences, organelle and other interspersed repeats were automatically removed by the software. The software used CAP3 [13] to assemble the sequences into contigs and singletons with the criterion of 80% overlap identity between one end to another end of read. The results were obtained in different output files e.g. contigs and singletons. For the purpose of SSR identification, the contig and singleton sequences were combined to form a data set of non-redundant sequences.

2.3 Mining of SSRs from Assembled ESTs

To detect SSRs in the EST sequences, we used SSR Identification Tool (SSRIT) [14] (http://www.gramene.org/db/markers/SSRtool) The sequence search for SSR markers conducted using SSRIT was carried out by setting the search parameters to identify at least five repeats of SSR motifs with a maximum of ten bp. The program takes a FASTA format sequence file as an input. The generated output file contains sequence name, number of SSRs in the sequence, SSR type, SSR motif, repeat number, sequence coordinates for SSR and the length of the sequence. The average distance between SSRs was calculated by dividing the total length of each region (calculated by multiplying the length of assembled ESTs by the proportion of each region) by the number of SSRs in the region.

2.4 SSR-EST Similarity Searches

BLASTx analysis SSR-EST sequences against the Swissprot/Uniprot protein database was performed and most significant matches were recorded.

2.5 SSR Protein Domain Analysis

The EST-SSRs were searched for functional protein domains using PROSITE (http://www. http://prosite.expasy.org/prosite.html)

2.6 Primer Designing and Marker Validation

The SSRs detected were used for primer designing and 14 primer pairs were custom synthesized and validated for their ability for amplification on a set of 20 cauliflower genotypes [15].

3. RESULTS AND DISCUSSION

3.1 Redundancy in EST Sequences

Results reveal that ESTs of *Brassica oleracea* varieties are abundant in microsatellites. ESTs, however, often represent partial and redundant cDNA sequences, therefore the reduction in redundancy is essential before analyzing them for SSRs.

After EST sequence assembly, out of the 2564 ESTs, 254 were successfully assembled into contigs and 1435 were designated as singletons as they showed no overlap with any other EST. After assembly, the whole dataset was reduced to 1689 sequences which showed 34.12% of data redundancy (Table 1) (ESM 1). 8.6%, 0%, 10% and 11.1%, of ESTs in cauliflower, cabbage, broccoli and brussels sprouts, respectively, formed contigs, whereas 65.2%, 100%, 54.9% and 11.1% of the ESTs in cauliflower, cabbage, broccoli and brussels sprouts, respectively, were unique (Table 1). A non redundant group of 1689 ESTs consisting of contigs and singletons was formed. 34.12% reduction in redundancy was found which implies that the number of ESTs had been reduced by a sizeable proportion prior to the SSR analysis. This indicates that there is high amount of overlapping in EST sequences belonging to the same genome.

3.2 Frequency Distribution of SSR-Ests

The total number of SSRs was found to be 107. It was observed that the frequency percentage of dinucleotide motifs was 68.22% and for trinucleotides, it was 31.78%. The results suggest that, after assembly, dinucleotide SSRs comprise the most common motifs in the analysed data set.

Table 1. Results of EST sequence assembly

Name of crop	Total number of EST sequences	Number of singletons	Number of contigs	Non redundant data set (singletons+ contigs)	% reduction in redundancy
Cauliflower	265	173	23	196	26.03%
Brussels sprouts	9	1	1	2	77.77%
Cabbage	5	5	0	5	0%
Broccoli	2285	1256	230	1486	34.96%
Total	2564	1435	254	1689	34.12%

3.3 Average Distance between the SSRs

The total length of SSR-ESTs in the data set was 49348 Kb with the average distance between SSRs being 0.4 Mb. The average distance between two dinucleotide and trinucleotide SSRs was 11.4 Mb and 24.5 Mb, respectively.

3.3.1 Statistics of SSRs physical distances

No. of sequences with SSRs = 100
Total length of SSR-ESTs = 49348Kb
Average length of SSR-ESTs = 49348/100= 493.48 Kb
Total number of sequences searched for SSRs= 1689
Average distance between two SSRs= (1689×493.48)/100 = 8334.87 Kb
Average distance between two di-SSRs= (1689 ×493.48)/73= 11417.64 Kb
Average distance between two tri-SSRs = (1689 ×493.48)/34 = 24514.34 Kb

3.4 Frequency Distribution of Various SSR Types

Among dinucleotide SSRs, the most frequent repeat was AG/GA at 42.47% and GT (2.74%) was the least frequent (Table 2). The most frequent trinucleotide motif was AAG/GAA/GGA/GAG/AGA at 32.35% and ATC and GCT were the least frequent at 2.94% (Table 2).

3.5 Nature of Trinucleotide SSR Encoded Amino Acids

Every trinucleotide motif encodes an amino acid which has putative roles in activity of protein molecules. Out of a total of 34 trinucleotides, 17% trinucleotides SSRs encoded lysine followed by 11% coding for both asparatic acid and glutamic acid (Table 3).

SSR encoded amino acids were classified into polar and non polar on the basis of their nature. Polar amino acids were more frequent than non polar amino acids (53%).

3.6 Blastx Analysis for Sequence Identification

BLASTx analysis was carried out for the 100 EST-SSR sequences. Based on this analysis, a putative function could be assigned to 82 of the sequences (82%), assuming a threshold value of 1E-5. The annotation results indicated that most of the EST-SSR sequences from cauliflower and broccoli showed high homology with known *A. thaliana* proteins (ESM 1).

3.7 Analysis of SSR–Protein Domains

100 SSR containing sequences were analyzed for protein domains through PROSITE. It is apparent, when studying protein sequence families, that some regions have been better conserved than others during evolution. These regions are generally important for the function of a protein and/or for the maintenance of its three-dimensional structure. By analyzing the constant and variable properties of such groups of similar sequences, it is possible to derive a signature for a protein family or domain, which distinguishes its members from all other unrelated proteins. The functional domains identified were 2Fe-2S ferredoxin binding domain, iron-sulphur binding domain, 4Fe-4S ferredoxin binding domain and Anaphylatoxin domain signature, VWFC domain, Thaumatin family profile, EGF-like domain, Thiolases active site, Tubulin subunits alpha, beta and gamma domains, C-terminal cystine knot and Integrins beta chain cysteine-rich domains. As a result 83 sequences (83%) containing protein domains were identified. The SSR-FDMs (functional domain markers) provide information that these genetic markers once transcribed have putative functions (ESM 1).

3.8 Primer Designing and Marker Validation

The detected SSRs were subjected to primer designing using the online interface of the PRIMER3 software. Out of the designed primer, 14 were custom synthesized and were used for marker validation on a set of 20 cauliflower genotypes. The primers gave scorable bands (Fig. 1.) and revealed 52% polymorphism among the tested cauliflower genotypes. The primers were also able to classify the genotypes into different groups based on the genotypic data [15].

4. DISCUSSION

Simple sequence repeat based markers are an important class of molecular markers due to their abundance, hyper variability, high polymorphism and transferability. Microsatellite distribution and frequency of nucleotide repeats reflect the underlying mutational processes, selection constraints as well as DNA repair mechanisms. SSRs have a major role in studying genetic variation in evolution studies due to their functional qualities [16,17].

EST databases can be used effectively the development of SSR markers, as they are associated with transcribed genes. This approach is both cost and time intensive once sufficient amounts of EST sequences are available. SSR markers developed from ESTs could be of great value in filling gaps in existing linkage maps, identifying markers linked to genes of interest and to develop markers for cross species studies. The use of EST derived SSR markers for genetic diversity studies is a novel tool that indicates variation in transcribed genes.

Table 2. Distribution of repeat motifs

S. No.	Repeat motif	Number	Frequency
	Dinucleotide repeats		
1.	AG/GA	18+13=31	42.47%
2.	AC/CA	2+4=6	8.21%
3.	CT/TC	12+12= 24	32.88%
4.	GT	2	2.74%
5.	AT/TA	4+6=10	13.60%
	TOTAL	73	
	Trinucleotide repeats		
6.	CTC/TCT/CTT/TTC	5+1+1+1=8	23.53%
7.	AAC/CAC	1+1=2	5.88%
8.	CAG/AGC	1+1=2	5.88%
9.	AAG/GAA/GGA/GAG/AGA	2+2+3+2+2=11	32.35%
10.	ATC	1	2.94%
11.	GAT/ATG	4+2=6	17.65%
12.	GTT/TGT	1+2=3	8.82%
13.	GCT	1	2.94%
	TOTAL	34	

Table 3. Nature of amino acids coded by the trinucleotide repeats

Amino acid	Number of coding trinucleotide repeats	Frequency of occurrence	Nature of amino acid
Leucine	6	17%	Non – polar
Serine	2	5.8%	Polar
Phenylalanine	1	2.9%	Non – polar
Asparagine	1	2.9%	Polar
Histidine	1	2.9%	Polar
Glutamine	1	2.9%	Polar
Aspartic acid	4	11.7%	Polar
Glutamic acid	4	11.7%	Polar
Glycine	3	8.8%	Polar
Arginine	2	5.8%	Polar
Isoleucine	1	2.9%	Non – polar
Methionine	2	5.8%	Non – polar
Valine	2	5.8%	Non – polar
Lysine	2	5.8%	Non – polar
Cysteine	2	5.8%	Non – polar

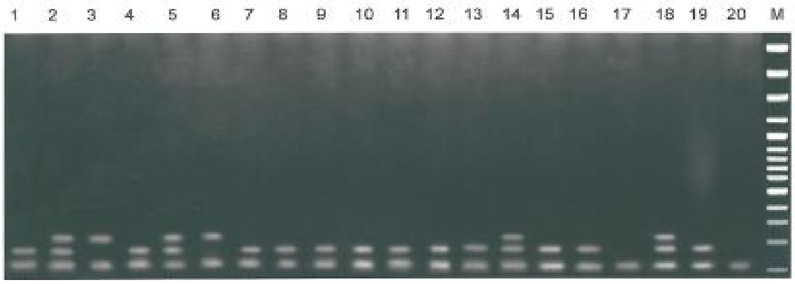

Fig. 1. Amplification profile of EST-SSR primer E5 with 20 cauliflower genotypes

Genotypes: 1: Shimla Clause, 2: Snowball 16, 3: Himalini Crystal Seeds, 4: US 178 Agri Seeds, 5: Cauliflower KT-22, 6: Girija Seminis, 7: Shoppers' Stock, 8: Snow Mystique Takii Seeds, 9: Hemant (Aghavi), 10: SCL 2003A Seedex Seeds, 11: Pallavi Crystal Seeds, 12: Madhavi Global Seeds, 13: Shubhra Seminis, 14: Shikha Krishi Samridhi Seeds, 15: Bonny Daenhfeldt, 16: Taj Nickerson Zwaan, 17: Himani Nunhams, 18: No. 71 Pahuja Seeds, 19: NU. 84 Pyramid Seeds, 20: Pusa Snowball PSBK 1

In the present study, the publicly available collection of 2564 ESTs from four different botanical varieties of *Brassica oleracea* (*B. oleracea* var. *botrytis*, *B. oleracea* var. *capitata*, *B. oleracea* var. *italica* and *B. oleracea* var. *gemmifera*) have been assembled using EGassembler web server. The server simplified the sequences into contigs and singletons using CAP3 assembly program. EST sequences were grouped into 1689 non-redundant EST sequences out of which 107 SSRs were detected in 100 ESTs (SSR-ESTs). Among all the SSR motifs the percentage frequency of dinucleotide SSRs was found to be maximum, the most frequent repeat being AG/GA and GT was detected to be the least frequent repeat. While among trinucleotide SSRs, the most frequent motif was AAG/GAA/GGA/GAG/AGA and ATC and GCT and GCT were the least frequent. Previous surveys carried out on microsatellite abundance analysis in plant genomes have shown ATT and CTT were the most frequent trinucleotide repeat motifs [18,19] and AT was the most abundant dinucleotide repeat motif followed by AG/CT and GT/CA [20,21]. He et al. [22] found that GA/CT repeat was the most frequently dispersed microsatellite in peanut. Cuc and coworkers indicated that GT/CA repeat motif was the most common, accounting for 37.6% of all repeat types, followed by GA/CT repeat at 25.9% [20]. From this it can be concluded that SSR motifs are not equally abundant in eukaryotic genomes and the relative abundance of different motifs varies among species [23,21].

The trinucleotide SSRs are triplet codon that code for a particular amino acid. It was observed that out of all the triplet codons, those encoding serine and lysine were predominant. The triplet codons form an open reading frame (ORF) translated to proteins. The analysis of data also revealed that the majority of the amino acids were polar in nature.

BLASTx analysis revealed strong homology of the sequences of cauliflower and broccoli to known *Arabidopsis* proteins. Also, the FDM analysis showed these sequences could code for a variety of protein domains. The sequences having both SSRs and FDMs indicate that the molecular markers were parts of sequences having some predicted function.

5. CONCLUSION

This study demonstrated that *in silico* mining of microsatellite loci is an efficient means for EST–SSR marker development. Microsatellites have proved very useful as molecular markers in many areas of genetic research including genome characterization and mapping. This study demonstrated the utility of computational approaches for detecting SSRs from publicly available plant EST sequences. Since EST–SSRs form part of transcribed regions of genome, therefore, they make a valuable resource both in structural as well as functional genomics in *Brassica oleracea* varieties and also in related members of Brassicaceae.

ACKNOWLEDGEMENTS

The research was funded by the University grant of the parent university.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Varshney RK, Thiel T, Stein N, Langridge P, Graner A. *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell Mol Biol Lett. 2002;7(2A):537–546.
2. Thiel T, Michalek W, Varshney RK, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet. 2003;106(3):411–422.
3. Tripathi KP, Roy S, Maheshwari N, Khan F, Meena A, Sharma A. SSR polymorphism in *Artemisia annua*: Recognition of hotspots for dynamics mutation. Plant Omics J. 2009;2(6):228-237.
4. Scott KD, Eggler P, Seaton G, Rosseto M, Ablett EM, Lee LS, Henry RJ. Analysis of SSRs derived from grape ESTs. Theor Appl Genet. 2000;100(5):723-726.
5. Cordeiro GM, Casu R, Mcintyre CL, Manners JM, Henry RJ. Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross-transferable to Helianthus and *sorghum*. Plant Sci. 2001;160(6):1115–1123.
6. Varshney RK, Graner A, Sorrels ME. Genic microsatellite markers in plants: Features and applications. Trend Biotechnol. 2005;23(1):48–55.
7. Morgante M, Hanafey M, Powell W. Microsatellites are preferentially associated with non-repetitive DNA in plant genomes. Nature Genet. 2002;30(2):194–200.
8. Toth G, Gaspari Z, Jurka J. Microsatellites in different eukaryotic genomes: Survey and analysis. Genome Res. 2000;10(7):967-981.
9. Rohrer GA, Fahrenkrug SC, Nonneman D, Tao D, Warre WC. Mapping microsatellite markers identified in porcine EST sequences. Animal Genet. 2002;33(5): 372–376.
10. Slate J, Hale MC, Birkhead TR. Simple sequence repeats in zebra finch (*Taeniopygia guttata*) expressed sequence tags: A new resource for evolutionary genetic studies of passerines. BMC Genomics. 2007;8:52.
11. Kantety RV, LaRota M, Matthews DE, Sorrells ME. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, *sorghum* and wheat. Plant Mol Biol. 2002;48(5-6):501-510.
12. Masoudi-Nejad A, Koichiro T, Shuichi K, Yuki M, Masanori S, Masumi I, Minoru K, Takashi E, Susumu G. EGassembler: Online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments. Nucleic Acids Res. 2006;34(suppl 2):459-462.
13. Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Res. 1999;9(9):868-877.
14. Temnykh S, Declerk G, Lukashova A, Lipovich L, Cartinhour S, Mccouch SR. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): Frequency, length variation, transposon associations and genetic marker potential. Genome Res. 2001;11(8):1441–1452.
15. Vaidya E, Kaur R, Bhardwaj SV. Data mining of ESTs to develop dbEST-SSRs for use in a polymorphism study of cauliflower (*Brassica oleracea* var. *botrytis*). J. Hortic Sci Biotech. 2012;87(1):57–63.
16. McCarthy JJ, Hilfiker R. Use of Single Nucleotide Polymorphism Maps in pharmacogenomics. Nature Biotechnol. 2000;18(5):505-508.
17. Pfost DR, Boyce-Jacino MT, Grant DM. A Snp Shot: pharmacogenetics and the future of drug therapy. Trends Biotechnol. 2000;18(8):334-338.
18. Lagercrantz ULF, Ellegren H, Andersson L. The Abundance Of various polymorphic microsatellite motifs differs between plants and vertebrates. Nucleic Acids Res. 1993;21(5):1111-1115.
19. Ferguson ME, Burow MD, Schulze SR, Bramel PJ, Paterson AH, Kresovich S, Mitchell S. Microsatellite identification and characterization in peanut (*Arachis hypogaea* L.). Theor Appl Genet. 2004;108(6):1064-1070.
20. Cuc LM, Mace ES, Crouch JH, Quang VD, Long TD, Varshney RK. Isolation and characterization of novel microsatellite markers and their application for diversity assessment in cultivated groundnut (*Arachis hypogaea*). BMC Plant Biol. 2008;8:55.

21. Han Z, Wang C, Song X, Guo W, Gou J, Li C, Chen X, Zhang T. Characteristics, development and mapping of *Gossypium hirsutum* derived EST-SSRs in allotetraploid cotton. Theor Appl Genet. 2006;112(3):430-439.

22. He G, Meng R, Newman M, Gao G, Pittman RN, Prakash CS. Microsatellites as DNA markers in cultivated peanut (*Arachis hypogaea* L.). BMC Plant Biol. 2003;3:3

23. Wang Z, Weber JL, Zhong G, Tanksley SD. Survey of plant short tandem DNA repeats. Theor Appl Genet. 1994;88(1):1-6.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history.php?iid=880&id=40&aid=9190