BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 11(3)

Mining, characterization and validation of EST derived microsatellites from the transcriptome database of *Allium sativum L*

Subodh Kumar Chand, Satyabrata Nanda, Ellojita Rout & Raj Kumar Joshi*

Functional Genomics laboratory, Centre of Biotechnology, Siksha O Anusandhan University, Bhubaneswar-751003, India; Raj Kumar Joshi – Email: rajkumar.joshi@yahoo.co.in; Phone: 09437684176; *Corresponding author

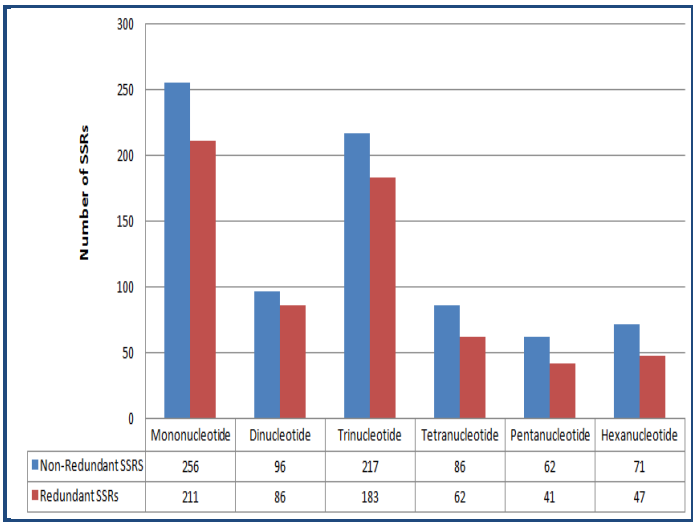
Received February 07, 2015; Accepted February 28, 2015; Published March 31, 2015

Abstract:
Expressed Sequence Tags (ESTs) with comprehensive transcript information are valuable resources for development of molecular markers as they are derived from conserved genic regions. The present study highlights the mining of EST database to deduce the class I hyper variable SSRs in *A. sativum*. From 21694 garlic EST sequences, 642 non-redundant SSRs were identified with an average frequency of 1 per 14.9 kb of garlic transcriptome. The most abundant SSR motifs were the mononucleotides (32.86%) followed by trinucleotides (28.50%) and dinucleotides (13.39%). Among the individual SSRs, (A/T)n accounted for the highest number (137; 21.33%) followed by (G/C)n (74; 11.52%) and (AAG)n (63;9.81%). Primers designed from a robust set of 7 AsEST-SSRs resulted in the amplification of 63 polymorphic alleles in 14 accessions of garlic. The resolving power of the markers varied from 4.286 (AsSSR7) to 18.143 (AsSSR13) while the average marker index (MI) was 5.087. These EST-SSRs markers for garlic could be useful for the improvement of garlic linkage map and could be used for evaluating genetic variation and comparative genomics studies in *Allium* species.

Keywords: *Allium sativum*, expressed sequences tags, EST-SSRs, SSR motifs

Background:
Garlic (*Allium sativum* L.), the "spice of life" is a unique monocot plant from the economically important family Liliaceae. It belongs to the genus *Allium* consisting of 1250 species which has been used throughout history for both culinary and medicinal purposes **[1].** It is attributed with several medicinal properties including being a stimulant, diaphoretic, an expectorant, a diuretic and a tonic due to the presence of allicin in the bulb and garlic oil in both bulb and leaves **[2].** Presently it is considered to be useful for the treatment and prevention of a number of diseases, including cancer, coronary heart disease, obesity, diabetes type-2, hypertension, cataract and disturbances of gastrointestinal tract **[3].** The interest and demand in garlic has significantly increased in the recent times due to its nutritional and pharmaceutical properties. India contributes about 4.1% of global garlic production and lies next only to China in terms of productivity **[4].** The Food and Agricultural organization of the United Nations has estimated an annual growth rate of 4.7% in the world demand for garlic. Although sexual reproduction is possible in garlic, nearly all garlic plants are propagated vegetatively. Clonal selection is the main breeding method for modern garlic since no sexual propagation in garlic usually precludes crop improvement through hybridization. In garlic breeding programme, genetic variation is increased only by somaclonal variation, genetic transformation and mutation **[5].** Continuous domestication of preferred genotypes together with their exclusive asexual propagation have eroded the genetic base of this crop and make the plant vulnerable to various biotic and abiotic stresses which together attributes for upto

60% yield losses **[4].** Therefore, it is highly essential to develop efficient and reliable molecular markers to analyse the genetic diversity among garlic species for conservation of germplasm and improvement of the extant crop. Several molecular marker systems are used in plants for characterization and mapping of important traits **[6].** Among them, the most reliable are the microsatellites.



	Mononucleotide	Dinucleotide	Trinucleotide	Tetranucleotide	Pentanucleotide	Hexanucleotide
Non-Redundant SSRS	256	96	217	86	62	71
Redundant SSRs	211	86	183	62	41	47

Figure 1: Reduction in redundancy and assembling *Allium sativum* ESTs by CAP3 program.

Microsatellites or simple sequence repeats (SSRs) are small array of tandemly arranged 1 to 6bp nucleotides present throughout the genome and mainly used as marker in genetic variation and population genetic study **[7].** SSRs have great advantages over other markers as they are simple, highly abundant, polymorphic, polyallelic, co-dominant and occur in both coding and non-coding regions of the genome **[8].** Although several SSR markers have been identified in garlic **[9, 10],** additional SSRs with polymorphism are needed, particularly for the development of linkage maps for use in trait specific mapping studies. In the recent years, high throughput next generation sequencing technologies has led to the generation of large databases of expressed sequence tags (ESTs) and genomic sequences. ESTs are the short and single pass sequence reads of mRNAs or cDNAs corresponding to the partial coding sequences of expressed genes and acts as an attractive alternative source for mining of SSRs from the coding regions of the genome **[11].** EST-SSRs are more advantageous than the genomic SSRs due to their easy availability and high transferability to related species thereby serving as reliable markers for gene mapping analyses **[12].** Use of ESTs to generate EST-derived SSRs has been reported in several plants species **[13].**

Considering the above facts into account, the present study aims to exploit the EST database of *Allium sativum* to develop EST derived SSRs. There are 21694 numbers of *Allium sativum* ESTs available in the dbEST database of National Centre for Biotechnology Information (NCBI) (as of 22nd December 2014). A user friendly SSR identification tool-SSRLocator **[14]** was used for this purpose. Further, primers were designed from a selected set of EST-SSRs to determine polymorphism and validate their utility by evaluating genetic diversity among different garlic accessions.

Methodology:

A total of 21,694 expressed sequence tag (EST) sequences were downloaded from the dbEST database hosted in GenBank at NCBI using the key word '*Allium sativum*'. The 21694 ESTs retrieved were from three different garlic tissues i.e leaf, stem and root. The Cross_Match program **[15]** was used with parameter set at minmatch ≥13 and minscore ≥20 to screen the ESTs against the UniVec database (ftp: //ftp.ncbi.nih.gov /pub/ UniVec/) to detect vector and adapter sequences. Additionally, the polyA/T tails and X characters were removed from the EST sequences using EST_trimmer.pl script (http: / / pgrc.ipk-gatersleben.de /misa /download /est_trimmer.pl). The trimming of the ESTs were done until no stretch of $(A/T)_5$ or $(X)_1$ was observed in a window of 100bp at the 5′ or 3′ end, respectively. The assembly program CAP3 **[16]** was used to assemble all the ESTs thereby creating a non-redundant dataset. The resulting output of unigenes, contigs and singlets were combined together as a selected non-redundant dataset for SSR identification. The SSR detection tool SSRLocator was used to detect EST-SSR loci from the garlic EST datasets. Two repeat motifs found close to each other within an EST were considered as individual entity and not compound SSRs as suggested by Gupta *et al.*, **[17].** Primers were designed from Class I SSR containing EST sequences using Primer3plus **[18]** and validated using NetPrimer (http: //www.premierbiosoft.com /netprimer /index.html). Primers having a score of more than 75 as evidenced by the absence of self dimer and/or cross dimer were selected and synthesized. 7 selected set of primers were tested for functionality and polymorphism against a panel of 14 *Allium sativum* accessions.

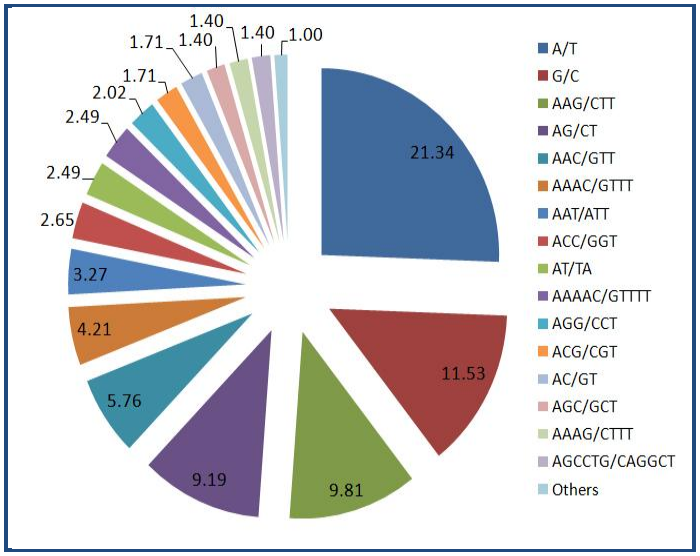


Figure 2: Distribution of EST-SSRs based on the motifs.

Results & Discussion:

The 21694 redundant EST sequences retrieved from NCBI represented approximately 11.87Mb of *Allium sativum* genome. During the scan for class I microsatellite repeats, 793 SSRs were detected in this dataset corresponding to 1.0 SSR per 14.9 kb. 53672 bp of empty vectors, low-quality sequences and Poly A/T tails were removed successfully during the pre-processing. Rest of the sequences were clustered and assembled into a non-redundant dataset of 14054 unique gene sequences (1491 contigs and 12563 singlets). Mining of Class I microsatellites revealed 642 unique SSR containing sequences within the non-

redundant datasets accounting for one SSR per 11.19 kb of garlic genome **Table 1 (see supplementary material).** The reduction in redundancy of SSR's obtained by trimming and clustering of non-redundant sequences is shown in **Figure 1.** Assembling of ESTs into contiguous or single gene sequences reduces the redundancy to calculate the accurate frequency and design unique primer sets. The parameters for SSR detection and frequency analysis are highly variable in different plant species **[17, 19].** Cradle *et al.,* **[20]** used a comprehensive computational strategy to estimate the average distances between SSRs in non-redundant ESTs of various plant species such as rice (3.4Kb), soybean (7.4Kb), tomato (11.1Kb), *Arabidopsis* (13.8Kb), poplar (14.0Kb) and cotton (20.0kb). We followed the same strategy and found one SSR per 11.19 kb of garlic non-redundant EST dataset. This suggests that frequency of EST-SSRs in the expressed portion of the garlic genome is high in comparison to rice, soybean and tomato and lower than other plant species.

The distribution of the individual SSR motifs among the non-redundant set of 642 SSRs is represented in **Table 2 (see supplementary material).** The mined EST-SSRs were classified as simple motif type, with a single motif; and compound SSRs, with more than two motifs. Among the 642 EST-SSRs, 12 loci (1.86%) represented compound SSRs while the rest 630 loci (98.14%) consisted of simple repeats **(Figure 2).** Among the Class I SSR motifs detected, mononucleotides were the most abundant (32.86%) followed by trinucleotides (28.50%) and dinucleotides (13.39%). The most common repeat motif was (A/T) in mononucleotides, (AG/CT) in dinucleotides, (AAG/CTT) in trinucleotides, (AAAC/GTTT) in tetranucleotides, (AAAAC/GTTTT) in pentanucleotides and (AGCCTG/CAGGCT) in the hexanucleotides. The high occurrence of mononucleotide repeats in the poly A/T trimmed dataset suggests that they are located within the expressive regions and not at the end of the mRNA sequences. This is in agreement with earlier report in *Catharanthus* **[21]** and turmeric **[22].** The dinucleotide motif AT (18%) had the lowest abundance while AG (68.6%) showed the highest frequency among the dimeric SSRs **Table 3 (see supplementary material).** The deficit of AT SSRs in EST sequences is in compliance with reports from rice **[23]** and *Arabidopsis* **[20].** AG/CT corresponds to GAG, AGA, UCU and CUC codons encoding for arginine, glutamic acid, alanine and leucine respectively. The surplus of AG/CT in the garlic genome corroborate with the fact that majority of plant proteins contains higher concentration of alanine and leucine. Among the trimeric SSRs, AAG (34.4%), AAC (20.2%) and AAT (11.4%) were the most common patterns in the garlic ESTs. AAG is the most common plant trinucleotide repeat as has been reported with highest percentage in cotton **[24]** and *Catharanthus* **[21].** AAC/GTT and AAT/ATT repeats are also significantly represented in periwinkle **[21],** turmeric **[22]**, pearl millet **[25]** and barley **[26].** Likewise, majority of the monocot genomes boast a specific occurrence of CGG repeats **[27].** ATT represents only 4% of the total garlic SSRs which is in accordance with the fact that most plants have the least percentage of ATT repeats because TAA-based variants encode stop codons [27]. The most frequent tetranucleotide SSR motifs were AAAC/GTTT (27) and AAAG/CTTT (09). The pentanucleotide, hexanucleotides repeats and the compound SSRs accounted for less than 3% contribution to the total SSR patterns.

Seven AsEST-SSRs primers designed from the selected set of SSR loci resulted in the amplification of 82 unambigously scorable bands in 14 accessions of *Allium sativum* **Table 4 (see supplementary material).** The test retrieved 63 polymorphic bands (76.8%), averaging 9 bands per primer. Of the 63 polymorphic loci, 39 (61.9%) were highly informative, as they were characterized by allele frequencies ranging from 0.2 to 0.8. The 7 AsEST-SSRs had a polymorphism information content (PIC) range of 0.689-0.780 with an average of 0.730. Thus, the markers exhibited good power to discriminate among the genotypes of *Allium sativum* used in this study. The resolving power of the markers varied from 4.286 (AsSSR7) to 18.143 (AsSSR13). The AsEST-SSRs were characterized by an average marker index (MI) of 5.087 with values ranging from 1.540 (AsSSR9) to 8.504 (AsSSR2). Three pairs of AsEST-SSRs (AsSSR2, AsSSR8 and AsSSR11) revealed high level of polymorphism among the garlic accessions by revealing higher number of alleles. Hence these EST derived SSR primers have potential as informative markers for genetic diversity analysis and selective trait identification studies in garlic improvement programmes.

Conclusion:

Microsatellites markers have myriad uses in plant genome analysis. The present study highlights the frequency, type and distribution of garlic EST derived microsatellites and demonstrates the successful development of EST-SSR markers in garlic accessions. Reproducible EST-SSR markers developed in this study could enrich the molecular marker resource for garlic and could be applied for trait mapping, assessment of genetic diversity, marker-assisted selection and functional analysis of candidate genes in commercial inbred varieties as well as in related *Allium* species.

Acknowledgement:

The authors gratefully acknowledge the research support from Science and Engineering Research Board (SERB), Dept.of Science and Technology, Govt.of India and thankful to Prof. Sanghamitra Nayak, Head, Centre of Biotechnology, Siksha O Anusandhan University for her encouragement and support.

References:

[1] Kik CKR & Gebhardt R, *Nutr Metab Cardiovasc Dis.* 2001 **11**: 57
[2] Banerjee SK *et al. Phytotherapy Res.* 2003 **17**: 97 [PMID: 12601669]
[3] Galano A & Marquez MJ, *Phys Chem.* 2009 **113**: 16077 [PMID: 19904959]
[4] https://www.faostat.fao.org/default.aspx/yearbook2013.
[5] Kenel F *et al. Plant Cell Rep* 2010 **29**: 223 [PMID: 20099065]
[6] Agrawal M *et al. Plant Cell Rep.* 2008 **27**: 617 [PMID: 18246355]
[7] Li X *et al. BMC Genet.* 2014 **15**: 124 [PMID: 25481640]
[8] Gupta PK & Varshney RK, *Euphytica* 2000 **113**: 163
[9] Cunha CP *et al. Am J Bot.* 2012 **99**: e17 [PMID: 22203654]
[10] Ipek M *et al. Sci Agricola* 2015 **72**: 41
[11] Davey JW *et al. Nat Rev Genet.* 2011 **12**: 499 [PMID: 21681211]
[12] Varshney RK *et al. Trends in Biotech.* 2005 **23**: 48 [PMID 15629858]
[13] Ellis JR & Burke JM, *Heredity* 2007 **99**: 125 [PMID: 17519965]
[14] Da Maia *et al. Int J Plant Genom.* 2008 **1**: 9 [PMID:18670612]

[15] Li W *et al. Bioinformatics* 2006 **22**: 1658 [PMID: 16731679]
[16] Huang X & Madan A, *Genome Res* 1999 **9**: 868 [PMID: 10508846]
[17] Gupta PK *et al. Mol Genet Genomics* 2003 **270**: 315 [PMID: 14508680]
[18] Untergasser L *et al. Nucleic Acids Res.* 2007 **35**: W71 [PMID: 17485472]
[19] Gao L *et al. Mol Breed* 2003 **12**: 235 [DOI: 10.1023/A:1026346121217]
[20] Cardle L *et al. Genetics* 2000 **156**: 847 [PMID: 11014830]
[21] Joshi *et al. Bioinformation* 2011 **5**: 378 [PMID: 21383904]
[22] Joshi RK *et al. Bioinformation* 2010 **5**: 128 [PMID: 21364792]
[23] Temnykh S *et al. Genome Res.* 2001 **11**: 1441
[24] Lu YD *et al. Chin Sci Bullet* 2008 **55**: 1889
[25] Senthilvel S *et al. BMC Plant Biol* 2008 **8**: 119 [PMID: 19038016]
[26] Thiel T *et al. Theor Appl Genet* 2003 **106**: 411 [PMID: 12589540]
[27] Chin ECL *et al. Genome* 1996 **39**: 866 [PMID: 8890517]

Edited by P Kangueane
Citation: **Chand *et al.*** Bioinformation 11(3): 145-150 (2015)

BIOINFORMATION

open access

Supplementary material:

Table 1: Summary of EST-derived microsatellites from the EST database of *Allium sativum.*

Parameters	Values
Total number of ESTs	21694
Total sequence analyzed	11866618bp
Total number of SSR identified including Poly A/T	793
Length of ESTs after Poly A/T tail removal	11812946bp
Total gene sequences after assembly	14054 (7184296bp)
Total number of contigs	1491 (2783241bp)
Total number of singletons	12563 (4401055bp)
Total number of unique Class I SSR loci located	642
Frequency of Class I SSRs in garlic ESTs	1 per 11.19 kb

Table 2: Distribution of SSR motifs in *Allium sativum.*

Repeat motif type	Number	Frequency (%)	Most abundant motif	Frequency within their own repeats (%)
Mononucleotide	211	32.86	A/T	64.92
Dinucleotide	86	13.39	AG/CT	68.60
Trinucleotide	183	28.50	AAG/CTT	34.42
Tetranucleotide	62	9.65	AAAC/GTTT	43.54
Pentanucleotide	41	6.38	AAAAC/GTTTT	39.02
Hexanucleotide	47	7.32	AGCCTG/CAGGCT	19.14
Compound	12	1.86	//	//
Total	**642**	**100**	//	//

Table 3: Total number of detected SSR loci

Motif	Number of loci	Motif	Number of loci
A/T	137	AAAAC/GTTTT	16
G/C	74	AAAAT/ATTTT	7
AT/TA	16	AAAGG/CCTTT	5
AC/GT	11	AATGG/CCATT	3
AG/CT	59	AAAAG/CTTTT	6
AAC/GTT	37	ACAGC/GCTGT	2
AAG/CTT	63	AAGAG/CTCTT	1
AAT/ATT	21	AGAGG/CCTCT	1
ACC/GGT	17	AGATCG/CGATCT	4
ACG/CGT	11	AAAATG/CATTTT	2
AGC/GCT	9	AACGCC/GGCGTT	5
AGG/CCT	13	AAGAGG/CCTCTT	3
ATC/GAT	7	ACCATC/GATGGT	2
CCG/CGG	5	AGCCTG/CAGGCT	9
AAAC/GTTT	27	AGCAGG/CCTGCT	6
AAAG/CTTT	9	ATCGGC/GCCGAT	1
AAAT/ATTT	7	AAAAAG/CTTTTT	4

AAGG/CCTT	4	ACCGCC/GGCGGT	3
AATT/TTAA	3	AAAAAC/GTTTTT	6
AGAT/ATCT	2	ACGGCG/CGCCGT	1
AGCC/GGCT	5	ACGAGG/CCTCGT	1
ATAG/TCTA	3	Compound SSRs	12
AATC/TGAT	2	**Total**	**642**

Table 4: Detection of polymorphism in14 accessions of *Allium sativum* by 7 selected EST derived SSR markers developed in this study

EST-SSRs	**Loci**			**PIC***	**Resolving power**	**Marker index**
	Amplified	**Polymorphic**	**% Polymorphism**			
AsSSR2	14	13	92.85	0.704	12.00	8.504
AsSSR3	15	11	73.33	0.754	17.28	6.087
AsSSR7	5	4	80.00	0.780	4.286	2.497
AsSSR8	13	11	84.61	0.708	14.571	6.596
AsSSR9	8	4	50.00	0.770	11.714	1.540
AsSSR11	13	10	76.92	0.710	15.429	5.463
AsSSR13	14	10	71.42	0.689	18.143	4.923
Total	82	63				
Mean	11.71	9		0.730		5.087

*Polymorphism Information content